Alexandre Leopoldo Gonçalves

UM MODELO DE DESCOBERTA DE CONHECIMENTO BASEADO NA CORRELAÇÃO DE ELEMENTOS TEXTUAIS E EXPANSÃO VETORIAL APLICADO À ENGENHARIA E GESTÃO DO CONHECIMENTO

Tese apresentada ao
Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do grau de Doutor em
Engenharia de Produção

Orientador: Prof. Roberto Carlos dos Santos Pacheco, Dr.

Florianópolis
2006

Alexandre Leopoldo Gonçalves

UM MODELO DE DESCOBERTA DE CONHECIMENTO BASEADO NA CORRELAÇÃO DE ELEMENTOS TEXTUAIS E EXPANSÃO VETORIAL APLICADO À ENGENHARIA E GESTÃO DO CONHECIMENTO

Esta tese foi julgada e aprovada para a obtenção do grau de **Doutor em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina

Florianópolis, 14 de dezembro de 2006.

Prof. Antônio Sérgio Coelho, Dr.
Coordenador do Programa

BANCA EXAMINADORA

Roberto C. S. Pacheco, Dr.
Universidade Federal de Santa Catarina
Orientador

Aran Bey Tcholakian Morales, Dr.
Universidade Federal de Santa Catarina

José Leomar Todesco, Dr.
Universidade Federal de Santa Catarina

Vinícius Medina Kern, Dr.
Universidade Federal de Santa Catarina

João José Vasco Peixoto Furtado, Dr.
Universidade de Fortaleza

Barend Mons, Dr.
Erasmus Medical Centre Rotterdam

Wesley Romão, Dr.
Universidade Estadual de Maringá

Aos meus pais, José e Izelda Gonçalves, e familiares pelo incentivo, pela confiança e pelos exemplos de força e dedicação.

Á minha esposa, por seu amor, compreensão e confiança nos momentos mais difíceis.

Ao meu filho, Victor, que iluminou nossa vida durante essa fase.

.

Agradecimentos

Ao Senhor, pela força, iluminação, saúde e imensa proteção. Por ter me colocado em contato com pessoas fantásticas e por me agraciar com a oportunidade de conhecer tantos amigos em um ambiente de conhecimento.

Gostaria de agradecer ao Programa de Pós-Graduação em Engenharia de Produção pela oportunidade em participar desse prestigiado curso. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento de parte dos estudos no *Knowledge Media Institute* (KMi) – *Open University*/UK, e ao Instituto Stela, onde foi possível a realização deste trabalho.

Agradecimento em especial ao meu orientador, professor Roberto C. S. Pacheco, e aos professores Aran B. T. Morales, José Leomar Todesco e Vinícius Kern, que contribuíram com opiniões e um apoio inestimável.

Obrigado aos amigos Dr. Enrico Motta, Dra. Victoria Uren e Dr. Jianhan Zhu, que contribuíram para a evolução e o amadurecimento do trabalho com suas críticas e sugestões.

Também gostaria de agradecer aos amigos que de alguma maneira participaram desta fase. Sou muito grato a Denilson e Graciele Sell, amigos e companheiros durante o período no exterior, Fabiano Beppler, pelas nossas discussões em assuntos relacionados ao presente trabalho, Carlos P. Niederauer, pelo apoio na fase inicial deste trabalho, Ricardo Rieke, Marcelo Domingos, Marlon Guérios, Wagner Igarashi, Alessandro Bovo, Andréa Bordin, pela oportunidade de trabalho conjunto, Sandra, Isabel e Paula, pela revisão deste trabalho, e a todos os amigos do Instituto Stela.

Finalmente, agradeço a todos que direta ou indiretamente contribuíram para a realização desta pesquisa.

.

"A sabedoria é a coisa principal; adquire pois a sabedoria, emprega tudo o que possuis na aquisição de entendimento".

Provérbios 4:7

Resumo

GONÇALVES, Alexandre Leopoldo. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento**. 2006. 196 f. Tese (Doutorado em Engenharia de Produção) ênfase em Inteligência Aplicada - Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

Atualmente as informações textuais, disponíveis nos mais variados formatos, constituem-se como um importante recurso uma vez que mapeiam parte considerável das atividades diárias nas organizações. Nesse sentido, os desafios residem em como sintetizar grandes volumes de informação e em como revelar, através de processos automáticos ou semi-automáticos, o conhecimento latente inerente aos documentos, objetivando auxiliar o estabelecimento de estratégias que promovam suporte aos gestores organizacionais. Para tal, o presente trabalho propõe um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e na expansão de unidades de análise chamado *Latent Relation Discovery* (LRD). O processo de correlação identifica, considerando-se um elemento textual de origem, o conjunto de elementos textuais mais relacionados. Esses relacionamentos são então utilizados na expansão de unidades de análise, ou seja, na redefinição do contexto de documentos. A avaliação do modelo é realizada em cinco cenários comparando-se LRD com outros métodos, entre eles, *Latent Semantic Indexing* (LSI), *Phi-squared, Mutual Information* e derivação deste, e *Z score*. No primeiro cenário o modelo proposto foi aplicado à recuperação de informação e, em seguida, à tarefa de agrupamento de documentos. Os demais cenários utilizaram informação provida por avaliadores humanos e por um mecanismo de busca tradicional para mensurar o grau de aderência entre os pares de elementos textuais e os métodos acima mencionados. Em todos os cenários, LRD apresentou melhores resultados em relação aos demais métodos. A principal contribuição do trabalho reside na definição de um modelo de correlação e expansão vetorial com o intuito de descobrir relacionamentos latentes entre elementos textuais, promover melhoramentos na representação de documentos e fornecer suporte a aplicações de Engenharia e Gestão do Conhecimento.

Palavras-chave: Descoberta de Conhecimento; Métodos de Correlação; *Latent Relation Discovery* (LRD); Engenharia do Conhecimento; Gestão do Conhecimento.

Abstract

GONÇALVES, Alexandre Leopoldo. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento**. 2006. 196 f. Tese (Doutorado em Engenharia de Produção) ênfase em Inteligência Aplicada - Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

Nowadays, textual information is an important resource once it maps daily organization activities. Thus, the challenge lies on how to synthesize the mass of information as well as how to reveal, through automatic or semi-automatic processes, latent knowledge embedded in documents in order to help on establishing strategies to support organization managers. To tackle such challenge, this work proposes a knowledge discovery model based on the correlation of textual elements and expansion of analysis units called *Latent Relation Discovery* (LRD). Taking into account a source textual element the correlation process identifies the most related textual elements. These relationships are used in the expansion process of analysis units, that is, in the redefinition of the context of documents. The model evaluation is carried out based on five scenarios comparing LRD with other methods, such as, *Latent Semantic Indexing* (LSI), *Phi-squared*, *Mutual Information* (MI), a MI derivation, and *Z score*. In the first and second scenarios, the proposed model was applied toward information retrieval and clustering tasks, respectively. The remaining scenarios are based on information provided by both human evaluators and a traditional search engine. So, it is applied in order to measure the adherence degree among pairs of textual elements and the correlation methods quoted above. For all scenarios, LRD obtained better results. The main contribution of this work lies on the definition of a correlation and expansion vector model aiming to discover latent relationships among textual elements, to promote improvements on document representation and to provide support toward Knowledge Engineering and Management applications.

Key words: Knowledge Discovery; Correlation methods; Latent Relation Discovery (LRD); Knowledge Engineering; Knowledge Management.

Lista de figuras

Lista de tabelas

Lista de siglas

AI	Artificial Intelligence (Inteligência Artificial)
ANN	Artificial Neural Networks (Redes Neurais Artificiais)
CIDE	Collaborative International Dictionary of English
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COSA	Concept Selection and Aggregation (Agregação e Seleção de Conceitos)
CP	Communities of Practice (Comunidades de Prática)
CVM	Context Vector Model (Modelo de Vetores de Contexto)
DM	Data Mining (Mineração de Dados)
ER	Entidades Relacionadas
IE	Information Extraction (Extração de Informação)
IR	Information Retrieval (Recuperação de Informação)
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em Bases de Dados)
KDT	Knowledge Discovery in Text (Descoberta de Conhecimento em Bases de Dados Textuais)
KE	Knowledge Engineering (Engenharia do Conhecimento)
KM	Knowledge Management (Gestão do Conhecimento)
LRD	Latent Relation Discovery (Descoberta de Relacionamentos Latentes)
LSI	Latent Semantic Indexing (Indexação Semântica Latente)
MI	Mutual Information (Informação Mútua)
NER	Named Entity Recognition (Reconhecimento de Entidades)
NLP	Natural Language Processing (Processamento de Linguagem Natural)
PE	Production Engineering (Engenharia de Produção)
SRC	Sistemas de Recomendação Colaborativa
SOM	Self-Organizing Maps (Mapas Auto-Organizáveis)
SVD	Singular Value Decomposition (Decomposição de Valores Singulares)
TF-IDF	Term Frequency / Inverted Document Frequency (Freqüência do Termo / Freqüência do Documento Invertida)
TM	Text Mining (Mineração de Textos)
VMI	Vechtomova's Mutual Information (Modelo de Informação Mútua proposto por Vechtomova et al., 2003)
VSM	Vector Space Model (Modelo de Espaço Vetorial)
WSD	Word Sense Disambiguation (Resolução de Sentidos Ambíguos)

Sumário

1 INTRODUÇÃO

A tecnologia ensinou uma lição à humanidade: nada é impossível.

Lewis Mumford

O foco do presente trabalho reside na extração e correlação de elementos textuais e na expansão de estruturas vetoriais, e no modo como essas tarefas promovem suporte a aplicações de Engenharia e Gestão do Conhecimento. No contexto do trabalho, elementos textuais são considerados padrões que podem ser extraídos e/ou classificados a partir de documentos. A extração desses padrões é obtida utilizando-se técnicas oriundas das áreas de Processamento de Linguagem Natural (NLP), Estatística, Extração de Informação (IE) e Recuperação de Informação (IR). Em etapa posterior, esses elementos são atribuídos a determinadas classes (por exemplo: organização, pessoa e projeto), servindo de base para o modelo proposto, que, através da correlação de elementos textuais e da expansão vetorial, procura revelar conhecimentos latentes[1] sobre elementos textuais e seus relacionamentos em coleções de documentos, como, por exemplo, quem conhece o quê e/ou quem, em quais projetos trabalha e com quais organizações interage. Nesse sentido, declara-se que a contribuição deste trabalho reside nas áreas de Engenharia e Gestão do Conhecimento, possuindo como tema principal um modelo de descoberta de conhecimento.

O restante deste capítulo objetiva: a) delinear a motivação do trabalho que está inserida no contexto da Engenharia e Gestão do Conhecimento; b) declarar os problemas e os pressupostos abordados no trabalho explicitando as idéias centrais do modelo assim como os critérios utilizados na avaliação; c) apresentar o objetivo geral e os objetivos específicos do trabalho; d) descrever a metodologia aplicada no desenvolvimento da pesquisa; e)

[1] No contexto do trabalho latente refere-se a algo potencialmente existente, mas não facilmente evidenciado.

contextualizar o trabalho na Engenharia de Produção (PE); f) delimitar o escopo; e g) apresentar a estrutura do trabalho.

1.1 MOTIVAÇÃO E CONTEXTO

A informação tem aumentado em volume e relevância no atual cenário organizacional, sendo imprescindível na tomada de decisão. Além disso, cita-se o fato de a informação estar cada vez mais se tornando facilmente acessível. Isso promove o desenvolvimento de áreas responsáveis pelo armazenamento, pela recuperação e, mais recentemente, pela transformação da informação em conhecimento.

Entretanto, conhecimento é definido como o que as pessoas conhecem e envolve processos mentais de compreensão, entendimento e aprendizagem, os quais residem na mente das pessoas (WILSON, 2002). Wilson ainda afirma que tudo aquilo que é utilizado para expressar o que se sabe somente pode ser realizado através de mensagens, portanto, não se constitui conhecimento, e sim informação; avaliação similar é apresentada por Alavi et al. (2001). De acordo com Davenport e Prusak (2000) conhecimento é um misto de experiências, valores, informação contextual, e percepção pessoal que fornece um sistema de avaliação e incorporação de novas experiências e informação. Origina-se e é aplicado na mente das pessoas. Nas organizações, está frequentemente embutido não somente em documentos ou repositórios, mas também nas rotinas, processos, práticas e normas.

Nota-se, portanto, um foco nos conhecimentos e nas habilidades dos colaboradores. Nesse sentido, a economia baseada em fatores tradicionais de produção tem-se movido em direção à economia baseada na informação e no conhecimento (DRUCKER, 1992; NONAKA; TAKEUCHI, 1995; DAVENPORT; PRUSAK, 1997). Destaca-se ainda que, de maneira cada vez mais intensa, as organizações consideram o conhecimento e as habilidades

de seus colaboradores como os seus recursos mais valiosos. Conforme ressaltam Schreiber et al. (2002, p. 2), "conhecimento é o recurso-chave nas empresas".

Nesse sentido, não o conhecimento em si, mas determinados ativos de conhecimento, tais como redes de relacionamento, competências, interesses dos colaboradores e projetos podem ser externalizados/explicitados em diferentes fontes de informação e identificados posteriormente de modo a auxiliar na tomada de decisão. O mapeamento desses ativos possui um papel importante no cenário organizacional, uma vez que cresce a demanda pela informação formatada, agregada e com foco nos seguintes pontos: (a) gestão por competências, considerando-se que o conhecimento e as habilidades dos colaboradores têm se tornado o recurso mais valioso para as organizações; (b) formação de equipes de projetos; (c) compartilhamento de conhecimento; (d) aumento da produtividade dos colaboradores; e (e) retorno de investimentos.

Muitas dessas demandas possuem como fonte primária bases de dados que objetivam a sumarização de competências, tais como a coleção de currículos da Plataforma Lattes[2] ou o diretório sobre competências de testemunhas especialistas (DOZIER et al., 2003). Podem igualmente ser documentos ordinários, tais como páginas *Web*, relatórios técnicos, *e-mails* e registros de comunicação instantânea, os quais refletem o dia-a-dia nas atividades dentro das organizações. Mais recentemente citam-se os ambientes colaborativos de disseminação de informação, os chamados *wikis*[3].

Visando lidar com essas demandas, áreas como extração e recuperação de informação e descoberta de conhecimento possuem um importante papel no suporte às aplicações de Engenharia e Gestão do Conhecimento. Essas áreas promovem uma estrutura geral para revelar conhecimento latente a partir de coleções de documentos e como esse conhecimento

[2] Disponível em: <http://lattes.cnpq.br>.

[3] "O termo **wiki** é utilizado para identificar um tipo específico de coleção de documentos em hipertexto ou o software colaborativo utilizado na criação da coleção de documentos". Um exemplo é a enciclopédia livre **Wikipédia** (disponível em: <www.wikipedia.org>).

pode auxiliar no entendimento das relações estabelecidas intra e interorganização.

Tradicionalmente, a descoberta de conhecimento utiliza-se de bases de dados estruturadas. Todavia, considerando-se fontes textuais como repositórios de conhecimento latente, alterações nesse modelo são requeridas. Isso acontece principalmente na fase inicial, com a identificação de elementos textuais relevantes, bem como durante a fase responsável por aplicar métodos especializados, denominada Mineração de Texto (TM). O modelo como um todo passa a ser chamado de Descoberta de Conhecimento em Textos (KDT).

A Mineração de Textos destina-se ao descobrimento de padrões em textos de linguagem natural que possam revelar conhecimento útil, ou seja, aplicável à tomada de decisão. Para esse fim, métodos provenientes das áreas de extração e recuperação de informação são usualmente utilizados na identificação de elementos textuais relevantes. Apesar de as abordagens tradicionais de KDT fornecerem algoritmos capazes de lidar com a extração de conhecimento, não está claro como essas abordagens conduzem os padrões revelados em direção às aplicações de Engenharia e Gestão do Conhecimento. Sob a perspectiva da descoberta de conhecimento, elementos textuais e seus relacionamentos podem revelar padrões latentes que, no presente trabalho, caracterizam-se como ponto central do modelo proposto.

A Engenharia do Conhecimento (KE) nasceu como ramo da Inteligência Artificial (IA) e pode ser definida como o processo de aquisição do conhecimento de especialistas e transporte desse conhecimento para a forma computacional, os chamados sistemas baseados em conhecimento. Mais recentemente, a Engenharia do Conhecimento vem sendo redefinida como uma atividade de modelagem de conhecimento. Modelos são utilizados para capturar características-chave de determinado sistema do mundo real, dividindo-os em partes menores que podem ser mais facilmente gerenciáveis e entendidas. Modelos são em geral associados com o domínio que representam (SAVOLAINEN et al., 1995). Um modelo é, portanto, “uma

simplificação da realidade" (BOOCH et al., 1999). A essência da modelagem de conhecimento reside na representação de sistemas em determinados níveis procurando-se abstrair questões de implementação e focar, em vez disso, nas competências dos sistemas (MOTTA, 2000).

Por sua vez, a Gestão do Conhecimento (KM) caracteriza-se como um conjunto de ações disciplinadas e sistemáticas de que uma determinada organização se utiliza para obter retorno a partir do conhecimento disponível (DAVENPORT; PRUSAK, 1997). Segundo Schreiber et al. (2002), a gestão do conhecimento é uma arquitetura para melhorar a infra-estrutura do conhecimento nas organizações com o objetivo de obter o conhecimento certo para a pessoa certa, no formato e no tempo certos.

Entretanto, a efetividade de sistemas de Gestão do Conhecimento requer, em princípio, o auxílio da tecnologia (MARWICK, 2001). Requerem ainda metodologias, arquiteturas e tecnologias capazes de tornar o conhecimento acionável de modo que promova algum tipo de vantagem competitiva. Entre as metodologias, CommomKADS é uma referência, oferecendo um conjunto de instrumentos voltado ao ciclo de gestão do conhecimento (SCHREIBER et al., 1994; SCHREIBER et al., 2002) e aplicável às arquiteturas organizacionais visando gerir os ativos de conhecimento.

Nesse sentido, tais metodologias, arquiteturas, tecnologias e ferramentas oferecem suporte às aplicações de Engenharia e Gestão do Conhecimento. Entretanto, definir quais artefatos são capazes de promover um adequado gerenciamento dos ativos de conhecimento para que se tornem um diferencial competitivo constitui-se em tarefa não trivial.

1.2 DECLARAÇÃO DOS PROBLEMAS E PRESSUPOSTOS DA TESE

O aumento da complexidade nas relações entre as organizações e seus colaboradores exige mecanismos capazes de armazenar tais relações e de auxiliar no entendimento do

contexto que as envolve. Documentos produzidos pela interação desses diversos agentes fornecem um reflexo de suas atividades, mantendo assim registrados elementos textuais e seus relacionamentos.

De modo a se lidar com tais relações, métodos e algoritmos capazes de auxiliar na Engenharia e Gestão do Conhecimento são requeridos. Assim, a determinação do peso[4] dos relacionamentos entre elementos textuais e o estabelecimento de estruturas mais adequadas à representação de documentos podem ser entendidos como ferramentas úteis a análise de determinado domínio de problema.

1.2.1 DEFINIÇÃO DO PROBLEMA

A partir do contexto acima mencionado os seguintes problemas são identificados:

- Como explicitar os elementos presentes em uma comunicação textual que auxiliem no entendimento das relações estabelecidas por meio da interação entre os diversos agentes participantes dessa comunicação?
- Como identificar e gerir um conjunto de técnicas que, quando integradas, auxiliam na descoberta e na gerência dos diversos níveis de relacionamentos entre os agentes envolvidos nos processos de compartilhamento, disseminação e gestão dos ativos de conhecimento?
- Como incrementar o contexto de representação de fontes de informação textuais de modo a auxiliar na descoberta de conhecimento latente sobre as relações entre elementos textuais?

[4] Força pela qual dois elementos textuais estão conectados, determinando assim a relevância do relacionamento.

1.2.2 PRESSUPOSTOS DA TESE

Considerando os problemas acima mencionados os seguintes pressupostos da tese são apresentados:

- O aumento da complexidade nas relações entre pessoas (agentes de conhecimento) e a informação compartilhada entre elas esconde metaconhecimento, que somente pode ser identificado quando os elementos participantes, ou seja, agentes, conteúdo e relações, são apresentados de maneira integrada.
- A efetividade de modelos de descoberta de conhecimento passa pela integração de diversas técnicas de engenharia do conhecimento e serve de base ao desenvolvimento de sistemas capazes de auxiliar no gerenciamento dos ativos de conhecimento organizacional.
- Modelos de espaço vetorial têm sido aplicados em áreas como recuperação e agrupamento de documentos. Entretanto, a representação vetorial tradicional desses documentos, ou de qualquer outra fonte de informação, limita a eficácia dessas aplicações.

1.2.3 PROPOSTA DA SOLUÇÃO

A solução proposta neste trabalho para os problemas apresentados anteriormente possui como base a integração de técnicas de extração e recuperação de informação e a descoberta de conhecimento voltada a aplicações de Engenharia e Gestão do Conhecimento. No contexto do trabalho não serão discutidas questões culturais e sociais das organizações, mas sim como a tecnologia da informação, através das técnicas de extração e recuperação de informação, e descoberta de conhecimento, pode prover meios para auxiliar na gerência dos ativos de conhecimento intra e interorganizações.

As técnicas de Extração de Informação (IE) possibilitam a identificação de elementos textuais (entidade, conceitos e termos) relevantes a partir de texto completo. Uma entidade é representada como um vetor *E*, composto de descrição (texto que identifica a entidade), classe e informações adicionais, ou seja, *E*={descrição, classe, <informações adicionais>}. As informações adicionais representam as posições onde a entidade ocorre no texto. Por sua vez, a Recuperação de Informação (IR) fornece o modelo de armazenamento de entidades e suas relações, bem como o ferramental para a recuperação dessa informação de maneira ágil. Sendo assim, a partir de uma entidade de origem, todas as entidades associadas são recuperadas, constituindo um vetor de conhecimento das relações que a entidade possui. Finalmente, o modelo faz uso da técnica de agrupamento, através da qual se torna possível o estabelecimento de relações indiretas entre as entidades. Essa abordagem tende a promover meios adicionais para o entendimento dos diversos modelos de redes, por exemplo, redes de colaboração e redes de competidores.

A seguir são declaradas as possíveis soluções para os três problemas apresentados neste trabalho.

Uma solução para o **problema 1** é prover um modelo capaz de estabelecer, de maneira eficiente, a força do relacionamento entre elementos textuais. A solução passa pela obtenção de informações disponíveis em documentos que possibilitem, de maneira eficiente, o cálculo das relações. Através da utilização de métricas advindas da área de recuperação de informação e funções de distância adequadas, propõe-se um modelo de correlação capaz de determinar com acurácia o grau de relacionamento entre elementos textuais.

Uma possível solução para o **problema 2** é estabelecer um modelo integrado voltado à descoberta de conhecimento latente. A proposta mescla técnicas das áreas tradicionais de extração e recuperação de informação, assim como técnicas de mineração de texto. Da extração de informação utiliza-se o reconhecimento de elementos textuais, enquanto a

recuperação de informação possibilita meios de armazenar e recuperar facilmente esses elementos e seus relacionamentos. Por sua vez, a mineração de textos, através de modelos de correlação e técnicas de agrupamento, permite o estabelecimento tanto de relacionamentos diretos quanto indiretos entre elementos textuais de modo a revelar conhecimento latente em coleções de documentos.

Para o **problema 3** uma solução seria a aplicação de métodos de correlação de elementos textuais que possibilitem a expansão do modelo vetorial de maneira eficiente e escalável. Através da expansão, novas dimensões são adicionadas à representação vetorial e o contexto de determinada unidade de análise é modificado. Isso permite tanto a recuperação de documentos que antes não atendiam a determinado critério de busca quanto o estabelecimento de maneira mais eficiente de agrupamentos de unidades de análise.

1.2.4 PROBLEMAS E LIMITAÇÕES RELACIONADOS À VALIDAÇÃO DO MODELO PROPOSTO

Apesar de existirem modelos/arquiteturas voltados a aplicações de Engenharia e Gestão de Conhecimento, eles nem sempre são integrados ou não estão facilmente disponíveis para avaliações e comparações. Além disso, a implementação e a disponibilização desse tipo de modelo/arquitetura não são uma tarefa trivial tornando difícil medir/avaliar a efetividade e a utilidade da arquitetura. Nesse sentido, a avaliação do modelo proposto neste trabalho concentra-se nos itens centrais, ou seja, no método de correlação e no modelo de expansão vetorial.

Destaca-se como problema principal a falta de conjuntos de dados padronizados específicos à avaliação de métodos de correlação de elementos textuais, ou seja, à avaliação do grau de relacionamento. A abordagem orientada à tarefa (discutida na Seção 4.1) possibilita, através de questionários distribuídos entre grupos de avaliadores, a determinação de um mapa inicial sobre a precisão do modelo. Entretanto, tal abordagem tende a introduzir

erros devido à subjetividade do processo, sendo necessários projetos cuidadosos para minimizar essa interferência.

Adicionalmente, menciona-se a dificuldade de medir/avaliar a efetividade das aplicações de Engenharia e Gestão de Conhecimento. De maneira geral, seriam adequados meios de mensurar o impacto que aplicações dessa natureza possuem sobre o contexto organizacional. Ambientes de cooperação como o Portal Inovação[5], por possuírem algumas características de aplicações de gestão do conhecimento, podem servir de base à avaliação. Nesse sentido, a incorporação de módulos de gestão do conhecimento, tais como redes sociais, e o monitoramento das atividades realizadas através desses módulos possibilitariam a determinação do impacto no estabelecimento das cooperações entre especialistas/grupos de pesquisa e empresas.

Finalmente, ressalta-se que os resultados dos experimentos obtidos através da utilização de determinados parâmetros, tais como configurações de janela e limiares, estão restritos aos conjuntos de dados considerados nas avaliações. Outra limitação refere-se à possibilidade de múltiplos sentidos para uma determinada entidade. Apesar de o modelo ser aplicável a qualquer coleção de documentos, a falta do tratamento de múltiplos sentidos de maneira automática ou não pode conduzir a erros no estabelecimento da força do relacionamento entre entidades.

1.2.5 MODELO DE AVALIAÇÃO

A avaliação utilizada no trabalho se concentra na validação do modelo de correlação de elementos textuais e expansão vetorial. Para tal, conjuntos de dados padronizados foram utilizados de modo a demonstrar a utilidade do modelo quando aplicado à recuperação de informação e ao agrupamento de documentos.

[5] <www.portalinovacao.info>

Com base nesses conjuntos de dados, cada documento contido na base é expandido utilizando o processo de correlação. Através da correlação, os elementos textuais mais relacionados são identificados e adicionados ao vetor que representa o documento, alterando o espaço vetorial original. Os vetores modificados são avaliados utilizando-se as medidas de precisão e lembrança para a tarefa de recuperação de informação, além de considerar o erro quadrático para a tarefa de agrupamento.

Adicionalmente, o método de correlação é avaliado comparando-se cada relacionamento (pares de elementos textuais) estabelecido através de julgamentos efetuados por avaliadores ou através de um mecanismo tradicional de recuperação de informação. Por meio de medida padronizada, objetiva-se mensurar o grau de correlação do método proposto confrontando os resultados com outros métodos de correlação.

1.3 OBJETIVOS DO TRABALHO

1.3.1 OBJETIVO GERAL

O objetivo principal do trabalho é desenvolver um modelo de correlação de elementos textuais e expansão de unidades de análises visando revelar relacionamentos latentes em bases de dados textuais, promovendo, assim, meios de se analisar e entender determinado domínio de problema.

1.3.2 OBJETIVOS ESPECÍFICOS

Visando-se atingir o objetivo principal, alguns objetivos específicos são requeridos, entre eles:

- propor um modelo capaz de estabelecer relacionamentos diretos e indiretos[6] entre elementos textuais (entidades ou conceitos). Seja uma entidade *E*1, todas as demais entidades co-ocorrendo com essa na coleção de documento têm as suas relações explicitadas;
- comparar o método de correlação proposto com outros métodos que possibilitem a identificação de relacionamentos latentes.
- utilizar métodos de descoberta de conhecimento para estabelecer relacionamentos indiretos entre elementos textuais;
- propor um modelo de validação capaz de medir de maneira padronizada o desempenho dos principais componentes do modelo, a correlação e a expansão vetorial; e
- demonstrar a utilidade do modelo em aplicações de Engenharia e Gestão de conhecimento, tais como manutenção de ontologias, comunidades de prática, gestão por competência e localização de especialistas.

1.4 CONTEXTUALIZAÇÃO NA ENGENHARIA DE PRODUÇÃO

O principal interesse da Engenharia de Produção (EP) pode ser caracterizado como a busca por aprimoramentos nos sistemas de produção (*American Institute of Industrial Engineering*).

Conforme apresentado pelo curso de Engenharia de Produção da Universidade Federal de Santa Catarina (UFSC) (PPGEP, 2006) e seguindo o modelo sugerido pela Associação Brasileira de Engenharia de Produção, compete à Engenharia de Produção:

> O projeto, a implantação, a operação, a melhoria e a manutenção de sistemas produtivos integrados de bens e serviços, envolvendo homens, materiais, tecnologia, informação e energia. Compete, ainda, especificar, prever e

[6] São relacionamentos diretos quando co-ocorrem no mesmo documento e indiretos quando, apesar de não co-ocorrerem no mesmo documento possuem elementos textuais em comum.

> avaliar os resultados obtidos destes sistemas para a sociedade e o meio ambiente, recorrendo a conhecimentos especializados da matemática, da física, das ciências humanas e sociais, conjuntamente com os princípios e métodos de análise e do projeto da engenharia.

Essa concepção evidencia como as áreas da Engenharia e Gestão do Conhecimento podem ser contextualizadas na área da Engenharia de Produção, em especial, no que tange à pesquisa aplicada e o desenvolvimento de sistemas de informação e conhecimento. Ainda no contexto da Engenharia e Gestão do Conhecimento, o termo "sistema" possui um significado mais amplo, podendo representar organizações, seres humanos ou mesmo agentes inteligentes (MOTTA, 2000).

De modo geral, a união entre os sistemas de produção e de conhecimento objetiva melhoramentos em níveis organizacionais e, portanto, na engenharia de produção organizacional. Possibilita também uma ligação com a engenharia e gestão do conhecimento, auxiliando organizações na determinação de mapas mais detalhados do seu capital social bem como de ferramentas para gerir e disseminar o conhecimento.

Desse modo, o presente trabalho, que envolve a proposta de um modelo visando o suporte às aplicações e à Engenharia e Gestão do Conhecimento, insere-se no contexto da Engenharia de Produção.

1.5 METODOLOGIA DA PESQUISA

Método é uma série codificada de passos executados para realizar uma tarefa ou alcançar um objetivo específico. Nesse sentido, métodos científicos são essenciais para a aquisição de conhecimento baseado em evidências físicas. Observações, hipóteses, deduções e proposições são levadas em conta para explicar fenômenos naturais como teorias. O método científico, portanto, constitui-se no caminho para construir um sólido entendimento da natureza. Por outro lado, a metodologia utiliza-se desse conjunto de métodos, regras e procedimentos aplicando-os em uma disciplina em particular.

Considerando-se a metodologia empregada neste trabalho para a realização e a validação do modelo proposto, alguns estágios são delineados (Figura 1) e descritos a seguir.

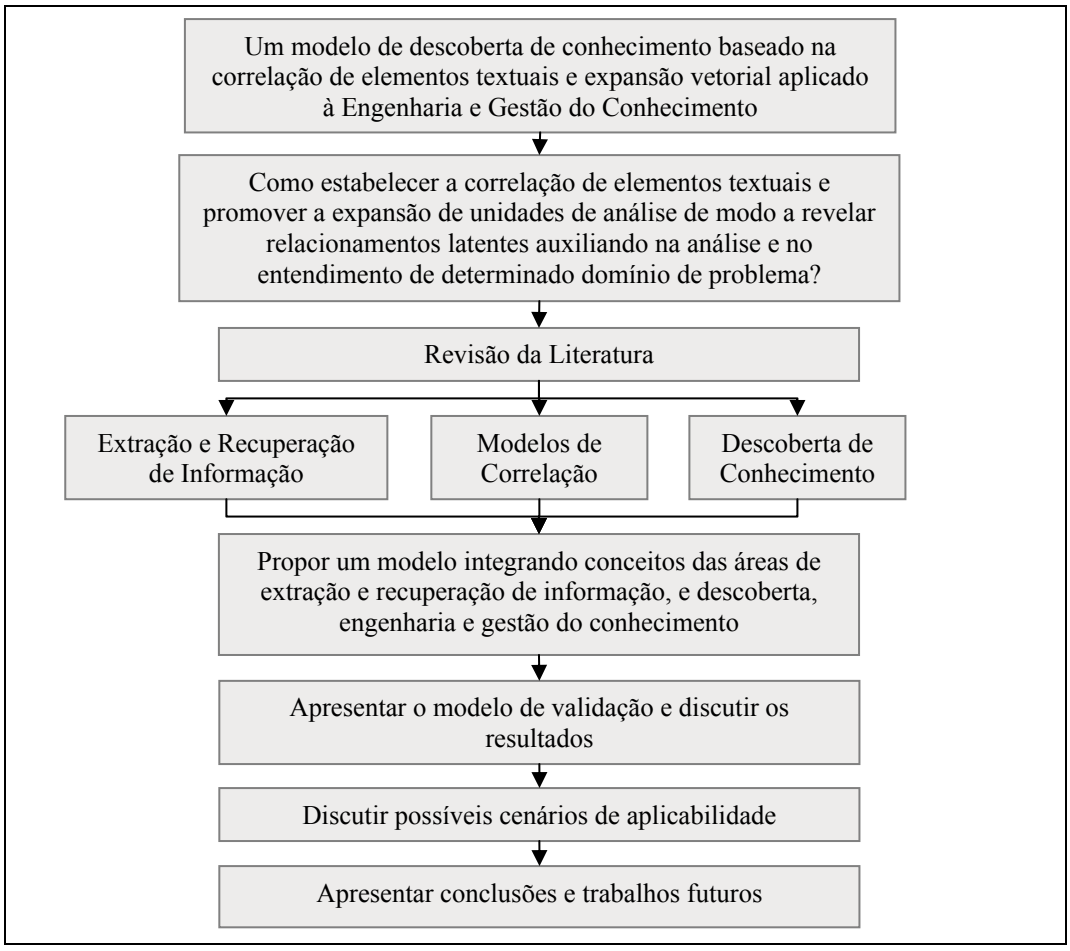


Figura 1 - Metodologia aplicada para o desenvolvimento do trabalho

- Revisão da literatura: nesta fase revisam-se as três principais áreas que promovem suporte ao desenvolvimento do trabalho: (a) Extração de Informação como a base para a aquisição de elementos textuais e a Recuperação de Informação para a representação de unidades de análise através do modelo de espaço vetorial; (b) Métodos de Correlação de elementos textuais como o núcleo do modelo, visando integrar ambas as áreas; e (c) Descoberta de Conhecimento como o arcabouço que suporta o modelo proposto nesse trabalho.

- Modelo proposto: o modelo proposto é apresentado detalhando-se todos os seus componentes e como esses atingem os objetivos do trabalho quando integrados. Envolve tanto a proposta do método de correlação quanto do modelo de expansão vetorial, chamado de *Latent Relation Discovery* (LRD).
- Validação: nesta fase o modelo proposto é validado considerando-se os dois principais componentes, a correlação de elementos textuais e a expansão vetorial. De modo a evitar a avaliação subjetiva, dois conjuntos de dados padronizados são utilizados, sendo um voltado à tarefa de recuperação de informação e o outro, à tarefa de agrupamento de documentos. Adicionalmente, estabeleceu-se um modelo de avaliação do grau de relacionamento entre entidades, comparando-se cada par de entidades com informações coletadas através de um mecanismo de busca tradicional e avaliações prévias fornecidas por especialistas. Em ambos os casos medidas padronizadas são utilizadas.
- Cenário de aplicabilidade: visando demonstrar a aplicabilidade do modelo, cenários de aplicação nas áreas de recuperação de informação e agrupamentos são apresentados. Além disso, discute-se a utilização do modelo de correlação voltado às aplicações de Engenharia do Conhecimento, tais como manutenção de ontologias e Gestão do Conhecimento, por exemplo, comunidades de prática, localização de especialistas e gestão por competência.
- Conclusões e trabalhos futuros: aqui são apresentadas as conclusões obtidas através do desenvolvimento deste trabalho assim como se discutem os trabalhos futuros, visando ao aprimoramento do modelo.

1.6 DELIMITAÇÃO DO TRABALHO

Este trabalho propõe um modelo baseado na correlação de elementos textuais e na expansão de espaços vetoriais voltados a aplicações de Engenharia e Gestão do Conhecimento. De modo geral, o modelo se preocupa com a fase de identificação e classificação de elementos textuais (entidades, conceitos e termos), passando pela correlação e expansão vetorial que auxiliam nas seguintes fases de: a) descoberta de conhecimento, e b) suporte ao desenvolvimento de aplicações que ajudem no entendimento do conhecimento organizacional.

Entretanto, está além do escopo avaliar possíveis melhorias no processo de descoberta, disseminação e compartilhamento de conhecimento, sejam essas obtidas através da implementação do modelo proposto ou através da utilização das aplicações de gestão do conhecimento.

Outro ponto a destacar refere-se ao estabelecimento das relações entre elementos textuais. O presente modelo objetiva o estabelecimento da força do relacionamento entre esses elementos, e não a identificação do tipo de relacionamento. Abordagens baseadas na anotação de textos e as centradas no usuário, sendo este responsável por gerir os relacionamentos obtidos a partir das anotações, são as mais utilizadas. Conforme afirmam Shadbolt et al. (2004), a realização desta tarefa de maneira automática é considerada um desafio para a área de extração de informação.

1.7 ORGANIZAÇÃO DA TESE

Este trabalho é composto de cinco capítulos, sendo os demais descritos a seguir.

- Capítulo 2. Revisão da Literatura: neste capítulo apresentam-se todas as técnicas consideradas no modelo proposto.

- Capítulo 3. Modelo Proposto: neste capítulo apresentam-se o método de correlação e expansão vetorial chamado LRD, assim como as demais fases que compõem o modelo.
- Capítulo 4. Apresentação dos Resultados: neste capítulo são demonstrados os resultados alcançados no desenvolvimento do trabalho.
- Capítulo 5. Conclusões e Trabalhos Futuros: neste capítulo são destacadas as conclusões sobre o trabalho realizado bem como são delineados os trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

Não há capital que dê melhores frutos a uma nação do que aquele que é posto à disposição dos jovens estudiosos e dos homens que, com inteligência, amor e liberdade, se dedicam à pesquisa científica.

Guilherme Guinle

2.1 INTRODUÇÃO

A "era da informação e conhecimento" tem apresentado novos desafios, tais como a implementação de técnicas computacionais capazes de revelar importantes padrões na crescente massa de dados. Isso se deve principalmente ao aumento da complexidade nas atividades organizacionais e na vertiginosa expansão da *Internet*. Cada vez mais se torna crucial que as organizações possuam um amplo conhecimento sobre as suas bases de dados textuais de modo que isso possa ser útil na tomada de decisão assim como possa prover melhoramentos nos processos e na gestão do capital intelectual. Nesse sentido, torna-se necessária a capacidade de extrair informações úteis, identificar relacionamentos e padrões latentes, e disponibilizar tais informações para análises em um contexto integrado. Considerando-se que parte importante das informações atualmente disponíveis nas organizações encontra-se na forma de texto, neste trabalho argumenta-se que documentos são a fonte primária para a descoberta de informações úteis à Engenharia e Gestão do Conhecimento. Os documentos que servem de insumo ao processo de extração de conhecimento podem ter como objetivo a sumarização de competências, tais como a coleção de currículos da Plataforma Lattes[7] ou o diretório sobre competências de testemunhas especialistas (DOZIER et al., 2003), mas eles podem igualmente ser documentos ordinários,

[7] Disponível em: <http://lattes.cnpq.br>.

tais como páginas *web*, relatórios técnicos, *e-mails* e registros de comunicação instantânea, que refletem o dia-a-dia das atividades dentro das organizações.

As próximas seções descrevem os principais conceitos de cada uma das áreas envolvidas na realização do trabalho, as quais promovem suporte a aplicações de gestão do conhecimento.

2.2 EXTRAÇÃO DE INFORMAÇÃO

A Extração de Informação (IE) caracteriza-se por métodos automáticos ou semi-automáticos voltados à identificação de fatos relevantes em coleções de documentos. Como afirma Freitag e Kushmerick (2000), IE caracteriza-se como um problema de conversão de informação textual (artigos, páginas *Web*, etc.) em objetos de dados estruturados adequados ao processamento automático de máquina. Conceito similar pode ser encontrado em Cardie (1997). Um dos primeiros sistemas voltados à extração de padrões foi o AUTOSLOG (RILOFF, 1993), que, através do analisador CIRCUS (LEHNERT et al., 1992), aprendia os padrões na forma de definições conceito-nodo para um domínio em particular.

Entretanto, o desafio reside na extração de informação relevante sem prévio aprendizado, uma vez que muitos dos algoritmos de Processamento de Linguagem Natural (NPL) requerem bases anotadas para treinamento (BONTCHEVA et al., 2004). A extração de informação tem se utilizado de bases anotadas, vocabulários controlados, recursos semânticos, tais como thesaurus, dicionários léxicos como *WordNet* (MILLER et al., 1990), ontologias (HEFLIN; HENDLER, 2000; VARGAS-VERA et al., 2002; ALANI et al., 2003b), métodos estatísticos (MANNING; SCHÜTZE, 1999) e aprendizado de máquina (CIRAVEGNA; WILKS, 2003).

Os dados a serem extraídos são em geral determinados por modelos que especificam listas de campos (*slots*) a serem preenchidos (NAHM; MOONEY, 2004; MANNING;

SCHÜTZE, 1999; CALIFF; MOONEY, 1999). Essa abordagem pode ser eficiente, mas a informação a ser extraída necessita ser anotada, geralmente por intervenção humana, o que se torna um processo custoso. Uma abordagem simplificada consiste na anotação de um pequeno número de textos com a informação a ser extraída. Através de métodos de aprendizado de máquina, a coleção de documentos rotulada é aplicada sobre grandes coleções de documentos (MOONEY; NAHM, 2005).

Recentemente, a combinação de técnicas vindas das áreas de Extração e Recuperação de Informação tem emergido, aproveitando-se de milhares de documentos já processados na *Internet* e nas organizações. Como exemplos que se utilizam dessa abordagem estão as arquiteturas WebFountain (GRUHL et al., 2004) e GATE (BONTCHEVA et al., 2004). Ambas fazem uso da Recuperação de Informação (IR) tentando primeiramente descobrir relevantes documentos em uma dada coleção de documentos, e a partir desses documentos utilizar-se de técnicas de IE para extrair informações relevantes.

A seguir serão discutidas abordagens automáticas utilizadas na extração de informação. Entre essas abordagens, citam-se a extração de entidades e a utilização de métodos estatísticos na identificação de seqüências naturais de palavras (apresentados na Seção 2.4).

2.2.1 EXTRAÇÃO DE ENTIDADES

A extração de entidades (*Named Entity Recognition* – NER) tem como objetivo descobrir nomes próprios, suas variações e classes (CUNNINGHAM, 2002; GROVER et al., 2002). Uma entidade pode ser definida como elementos textuais que representam objetos do mundo físico ou abstrato, como, por exemplo, “Albert Einstein” classificado como uma pessoa ou “Petrobrás” classificada como uma organização. Como declarado anteriormente, formalmente uma entidade é definida como um vetor *E* composto de uma descrição (a

estrutura textual), uma classe e informações adicionais, ou seja, *E*={descrição,classe,<informações adicionais>}. As informações adicionais podem indicar, por exemplo, um vetor de deslocamentos do padrão no documento. Gate (BONTCHEVA et al., 2004) e ESpotter (ZHU et al., 2005b) são exemplos de arquitetura e sistema utilizados na extração e na nomeação de entidades. Citam-se ainda os métodos DIPRE (BRIN, 1998) e Snowball (AGICHTEIN; GRAVANO, 2000), o sistema KNOWITALL (ETZIONI et al., 2004) e os trabalhos de Soderland (1999) e Ciravegna (2001).

O processo de NER depende basicamente de dois componentes - estruturas léxicas e padrões gramaticais - para identificar entidades e seus tipos. As estruturas léxicas são responsáveis pelo suporte ao processo formando a base de conhecimento (GUTHRIE, 1996). Cada classe a ser considerada durante a fase de extração (ex.: pessoa, organização, área de pesquisa) possui uma tabela léxica associada. Cada tabela léxica armazena as palavras referentes à sua classe, por exemplo, a tabela léxica da classe "Pessoa" teria nomes como "João", "Silva". Padrões são exibidos com regularidade na linguagem humana, por exemplo, em "as cidades de Florianópolis, São Paulo", "as <tipo> de <lista de instâncias>" constitui-se em um padrão da língua portuguesa. "Florianópolis" e "São Paulo" são entidades do tipo cidade; uma entidade que possua o padrão "Instituto <Instância>" é do tipo "Organização" ou "Instituto". Outro exemplo seria "<Instância >, Dr." indicando um padrão da classe "Pessoa". Existem outras informações relevantes ao processo que comumente podem ser utilizadas e que representam padrões, tais como endereços de *e-mails*, endereços de *internet*, datas, etc. Nesse sentido, a utilização de expressões regulares constitui-se em ferramenta útil na identificação de padrões como os acima listados.

Apesar de o processo de NER estar fortemente suportado por estruturas léxicas previamente construídas, um dos problemas observados refere-se aos múltiplos significados de uma entidade, ou seja, uma entidade pode pertencer a mais de uma classe. Assim, técnicas

de resolução de sentidos ambíguos (WSD) são requeridas. De modo geral, WSD envolve a associação de uma determinada palavra em um texto ou discurso com a definição ou o significado (sentido) que é distinto de outros significados potencialmente atribuídos para a mesma palavra (IDE; VÉRONIS, 1998; MANNING; SCHÜTZE, 1999). Cita-se como exemplo a palavra "jaguar", que pode ser, entre outras coisas, um animal ou a marca de um automóvel. Apesar de sua relevância no processo de extração de informação, mais especificamente, na extração de entidades, uma discussão mais detalhada sobre WSD está fora do escopo do trabalho.

2.2.2 MODELOS ESTATÍSTICOS

Modelos estatísticos têm sido amplamente utilizados na extração de informação, principalmente na identificação de "*collocations*"[8], seqüências de palavras que possivelmente representam conceitos a serem extraídos a partir de informação escrita. "C*ollocations*" são definidas como expressões compostas de duas ou mais palavras que correspondem a uma maneira convencional de dizer algo, ou seja, uma concatenação natural de palavras (MANNING; SCHÜTZE, 1999; SMADJA, 1993). Entre os métodos destacam-se o teste *t*, *Chi-square* (χ^2), *Phi-squared* (ϕ^2) (CONRAD; UTT, 1994; CHURCH; GALE, 1991), *Z score* (MANNING; SCHÜTZE, 1999; VECHTOMOVA et al., 2003), Informação Mútua (MI) (CHURCH; HANKS, 1990) ou derivação desse modelo desenvolvido por Vechtomova et al. (2003) e neste trabalho designado por VMI. Modelos com base mais empírica também têm sido aplicados, como, por exemplo, o algoritmo CORDER (ZHU et al., 2005a) e *Latent*

[8] Em Lingüística, uma combinação de palavras relacionadas dentro de uma sentença que ocorrem mais freqüentemente do que seria possível predizer em um arranjo aleatório de palavras; uma combinação de palavras que ocorrem com freqüência suficiente para serem reconhecidas como uma combinação comum, especialmente um par de palavras em que essas palavras ocorrem de maneira adjacente uma a outra (*Collaborative International Dictionary of English*, CIDE).

Relation Discovery (LRD) (GONÇALVES et al., 2006b). O detalhamento desses métodos será apresentado na Seção 2.4.

2.3 RECUPERAÇÃO DE INFORMAÇÃO

Meios de armazenamento e técnicas de indexação e recuperação da informação têm sido propostos constantemente nessa área. Embora atualmente existam diversas abordagens e ferramentas, o desenvolvimento de sistemas com tal propósito caracteriza-se como uma tarefa complexa e especializada. Isso se deve principalmente ao aumento do volume de informação assim como à disponibilidade nos mais variados formatos. Soma-se a isso a necessidade de a informação estar disponível de maneira fácil e rápida, e que seja útil ao usuário. Como Mitra e Chaudhuri (2000) afirmam, a explosão de informação resulta em grande demanda por meios eficientes e eficazes de organização, indexação e recuperação dessa informação.

De acordo com Baeza-Yates e Ribeiro-Neto (1999), recuperação, representação, armazenamento, organização e acesso são os principais processos na gestão da informação. Tais processos devem ser considerados de modo a prover aos usuários a recuperação da informação desejada.

Nesse sentido a Recuperação de Informação (IR) tem como tarefa básica possibilitar a localização de documentos que satisfaçam determinada consulta executada pelo usuário. Todavia, como Mitra e Chaudhuri (2000) sugerem, essa tarefa tem alguns desafios: (a) a natureza do tipo de informação não é estruturada; (b) documentos são escritos em linguagem natural; e (c) documentos cobrem uma ampla variedade de assuntos.

Apesar de essa ser uma área consolidada, novas abordagens têm sido desenvolvidas voltadas ao melhoramento do contexto semântico de documentos. Sendo assim algumas fases são requeridas. Primeiramente, encontram-se a fase de identificação de elementos textuais (termos, conceitos e entidades) relevantes a partir de documentos bem como a sua correta

normalização de modo a identificar sua importância em cada documento. Durante essa fase, dicionários controlados, taxonomias, thesaurus e extração de entidades podem ser utilizados para facilitar os processos de indexação e recuperação. Esse conjunto de elementos pode ser representado através de modelos, tais como o Modelo de Espaço Vetorial (VSM) (SALTON, 1975), o Modelo de Vetores de Contexto (CVM) (BILLHARDT et al., 2002) e o Modelo de Indexação Semântica Latente (LSI) (DEERWESTER et al., 1990). Através dessas representações, alguns modelos de consultas podem ser aplicados, entre os quais os modelos lógico (utilizando operadores AND, OR e NOT), vetorial, difuso e probabilístico. Em seguida, destaca-se a correlação de elementos textuais ou a utilização de ontologias e dicionários léxicos, tais como *WordNet*, com o objetivo de melhorar a representação de documentos e aumentar a precisão na recuperação de documentos. Por último, está a fase de disponibilização dessa informação de modo que usuários possam localizar facilmente o que desejam.

2.3.1 REPRESENTAÇÃO VETORIAL

O Modelo de Espaço Vetorial (VSM) é um dos modelos mais utilizados em aplicações de IR (MANNING; SCHÜTZE, 1999). Isso se deve à sua simplicidade conceitual e também porque esse modelo trata proximidade semântica como proximidade espacial. No VSM cada lista de termos (originada de documentos ou consultas) é considerada como um vetor de espaço *n*-dimensional, onde *n* é o número de distintos termos (RUSSEL; NORVIG, 1995). O conjunto de vetores forma a matriz termo–documento, armazenada, por exemplo, como uma estrutura de índice invertido.

Cada elemento pertencente ao espaço vetorial recebe um peso representando sua relevância no documento do qual foi extraído. Entre as equações de normalização, *tf-idf* (freqüência do termo pelo número de documentos nos quais o termo ocorre, *term frequency /*

inverted document frequency) é amplamente aplicada. De modo a recuperar documentos, faz-se necessário calcular a distância entre o vetor de consulta (termos informados pelo usuário) e os vetores que representam os documentos. A distância entre os vetores é calculada através de medidas de similaridade, por exemplo, produto interno (SALTON; BUCKLEY, 1988), co-seno (JONES; FURNAS, 1987), modelos probabilísticos ou difusos (KORFHAGE, 1997). Segundo Korfhage (1997), quando o modelo vetorial é utilizado, a medida de similaridade pode ser associada com a noção de distância, por meio da qual documentos que se encontram próximos no espaço vetorial são altamente similares ou com uma medida angular baseada na idéia de que documentos na mesma direção estão relacionados. O modelo vetorial é considerado flexível, pois facilmente possibilita que documentos recuperados possam ser classificados e avaliados de acordo com a sua relevância (NOUALI; BLACHE, 2003).

a) SIMILARIDADE DE VETORES

A similaridade entre termos é essencial no processo de recuperação de informação (KORFHAGE, 1997). O estabelecimento da similaridade ocorre entre documentos e consultas submetidas à base de documentos, e, dependendo da abordagem, a recuperação pode ser exata ou aproximada. Entre os modelos de comparação destacam-se: (a) modelo lógico; (b) modelo vetorial; (c) probabilístico; (d) difuso; e (e) baseado em proximidade (KORFHAGE, 1997).

Uma medida de similaridade denota o grau de similaridade (distância) entre conjuntos contidos em um universo Ω, onde Ω são usualmente coleções de documentos. Egghe e Michel (2002) discutem um conjunto de equações utilizadas na determinação de similaridade, entre elas, índice Jaccard, índice Dice, medida overlap (máxima e mínima), medida do co-seno e medida do pseudo-co-seno. Uma discussão ampla sobre medidas de similaridade é também apresentada por Jones e Furnas (1987).

Entre essas equações, o co-seno tem sido extensivamente aplicado a sistemas de recuperação de informação (SALTON; BUCKLEY, 1988). A equação do co-seno mede o ângulo entre dois vetores, variando de 1.0 (cos(0°) = 1.0) para vetores apontando na mesma direção, 0.0 (cos(90°) = 0.0) para vetores ortogonais[9] e -1.0 (cos(180°) = -1.0) para vetores apontando em direções opostas, sendo definido como:

$$\cos\theta = \frac{\sum_{i=1}^{n}(t_i \times q_i)}{\sqrt{\sum_{k=1}^{n}(t_k)^2} \times \sqrt{\sum_{j=1}^{n}(q_j)^2}} \quad (1)$$

onde t_i e t_k são as freqüências normalizadas dos ith e kth termos do vetor t, e q_i e q_j são as freqüências dos ith e jth termos do vetor q. Através dessa medida o ângulo entre as representações vetoriais pode ser calculado, ou através da comparação entre documentos e consultas, ou ainda através da comparação somente entre documentos.

2.3.2 INDEXAÇÃO SEMÂNTICA LATENTE

Indexação Semântica Latente (LSI) é definida como uma técnica automática que analisa as co-ocorrências de termos em bases textuais de modo a descobrir relacionamentos latentes entre eles (DEERWESTER et al., 1990). LSI tem promovido novos desenvolvimentos na área de recuperação de informação e tem sido aplicada em outras áreas. Destacam-se os modelos de tradução de textos, os modelos de mapeamento de funções cognitivas humanas (LANDAUER; DUMAIS, 1997), o mapeamento do gene humano, a recuperação de imagens, a classificação e/ou qualificação automática de textos, por exemplo, na avaliação de composições textuais. Todavia, LSI requer elevado tempo de processamento quando aplicada a grandes coleções de documentos (IKEHARA et al., 2001).

De modo a identificar as relações semânticas, LSI utiliza-se do modelo de Decomposição de Valores Singulares (SVD) (FORSYTHE et al., 1977). Considerando-se

[9] Que formam ângulos retos.

uma matriz esparsa termo–documento, a decomposição é calculada visando produzir uma matriz completa. Uma característica importante do modelo é a capacidade de redução de dimensionalidade através da utilização dos *k* fatores (valores singulares) mais relevantes. Esses fatores permitem recuperar parcialmente a informação que representa a matriz original. Tal abordagem assemelha-se à decomposição de autovetores[10] e à análise espectral e fatorial (DEERWESTER et al., 1990). Formalmente, qualquer matriz retangular *X* (representada como uma matriz termo–documento) pode ser decomposta como o produto de três outras matrizes:

$$X = T_o S_o D_0^{'}, \tag{2}$$

onde T_o e D_o possuem colunas ortonormais[11] e representam as matrizes de termos e documentos, respectivamente, e S_o representa a matriz diagonal de valores singulares.

Com o objetivo de exemplificar o método, considere a matriz *X* apresentada na Tabela 1. Nessa matriz cada termo está relacionado a um conjunto de documentos. Por questões de simplificação, as células da matriz são preenchidas como 0 e 1, indicando a existência ou não do termo para o documento. Entretanto, qualquer método de normalização de termos pode ser aplicado, por exemplo, o *tf-idf* mencionado anteriormente.

Termo	***D*1**	***D*2**	***D*3**	***D*4**	***D*5**	***D*6**	***D*7**	***D*8**
Descoberta	1	1	0	0	1	0	0	1
Conhecimento	1	0	0	0	0	1	0	1
Mineração	1	0	1	0	1	0	1	0
Interface	0	1	0	0	0	1	0	0
Usuário	0	1	0	0	1	1	0	0
Retroalimentação	0	1	1	0	1	1	0	0
Radiação	0	0	0	0	1	1	0	0
Turbina	0	0	0	0	1	0	1	0
Carvão	0	0	0	0	1	0	1	1

Tabela 1 - Exemplo de uma matriz termo–documento

[10] Na matemática, autovetor de uma transformação é um vetor que na transformação é multiplicado por um fator, chamado de autovalor do vetor em questão.

[11] Na álgebra linear, dois vetores *x* e *y* em um espaço de produto interno são considerados ortonormais se esses são ortogonais e de tamanho unitário, ou seja, a norma Euclidiana é 1.
Dois vetores, *x* e *y* em um espaço de produto interno *V* são ortogonais se o produto interno desses vetores <*x*, *y*> é zero.

Aplicando-se SVD, a matriz *X* é decomposta em três outras matrizes, *T*, *S* e *D*, respectivamente, satisfazendo a Equação 2. As tabelas abaixo (Tabela 2, Tabela 3, Tabela 4) demonstram o resultado do cálculo através de SVD.

0,4183	-0,0967	0,4886	0,0303	-0,6454	-0,3047	-0,0774	-0,2118
0,2543	0,0700	0,6826	-0,1070	0,5732	0,0965	0,0032	-0,0552
0,3673	-0,4767	-0,1678	-0,6517	0,1171	0,0337	-0,2343	0,2670
0,2154	0,4514	0,0193	0,0330	-0,0145	0,6105	-0,3730	-0,1566
0,3879	0,3457	-0,1693	0,2239	-0,1066	-0,1262	-0,2801	0,6355
0,4452	0,3145	-0,3155	-0,3057	-0,0855	0,0995	0,6144	-0,2670
0,2844	0,1505	-0,1763	0,2341	0,4638	-0,5991	-0,0069	-0,1566
0,2366	-0,3487	-0,3251	0,2808	0,0659	0,1037	-0,4027	-0,5341
0,3056	-0,4381	0,0417	0,5334	0,0531	0,3585	0,4265	0,2670

Tabela 2 - Matriz *T* (9x8) representando os vetores singulares esquerdos (dimensão de termos)

3,7653	0	0	0	0	0	0	0
0	2,2802	0	0	0	0	0	0
0	0	1,8184	0	0	0	0	0
0	0	0	1,3446	0	0	0	0
0	0	0	0	1,2221	0	0	0
0	0	0	0	0	0,7681	0	0
0	0	0	0	0	0	0,6518	0
0	0	0	0	0	0	0	0,000..

Tabela 3 - Matriz *S* (8x8) representando os valores singulares

0,2762	-0,2208	0,5518	-0,5417	0,0367	-0,2271	-0,4733	0,0000..
0,3896	0,4451	0,0127	-0,0138	-0,6972	0,3633	-0,1781	0,0000..
0,2158	-0,0711	-0,2658	-0,7121	0,0258	0,1733	0,5831	0,0000..
0,0000..	0,0000..	0,0000..	0,0000..	0,0000..	0,0000..	0,0000..	1,0000..
0,6495	-0,2410	-0,3430	0,2567	-0,1126	-0,5659	0,0605	0,0000..
0,4216	0,5842	0,0224	0,0581	0,6795	0,1056	-0,0650	0,0000..
0,2416	-0,5541	-0,2482	0,1209	0,1932	0,6455	-0,3230	0,0000..
0,2598	-0,2039	0,6670	0,3397	-0,0157	0,1957	0,5404	0,0000..

Tabela 4 - Matriz *D* (8x8) representando os vetores singulares direitos (dimensão de documentos)

Essas matrizes tornam possível a projeção de novos subespaços com o intuito de promover a redução de dimensionalidade. Sejam os valores singulares (*S*) classificados pela sua relevância, os primeiros *k* mais importantes fatores são considerados visando à recuperação aproximada do conteúdo original da matriz *X* de tal modo que:

$$X = TSD' \cong T_k S_k D_k^T, \tag{3}$$

Em geral valores de *k* entre 100 e 300 são considerados. No exemplo discutido aqui o valor de *k* é igual a 2. Com essa configuração as matrizes *T*, *S* e *D* terão suas dimensionalidades reduzidas, conforme apresentado a seguir (Tabela 5).

(a)	
0,4183	-0,0967
0,2543	0,0700
0,3673	-0,4767
0,2154	0,4514
0,3879	0,3457
0,4452	0,3145
0,2844	0,1505
0,2366	-0,3487
0,3056	-0,4381

(b)	
3,7653	0
0	2,2802

(c)	
0,2762	-0,2208
0,3896	0,4451
0,2158	-0,0711
0,0000	0,0000
0,6495	-0,2410
0,4216	0,5842
0,2416	-0,5541
0,2598	-0,2039

Tabela 5 - Matrizes *T*, *S* e *D* considerando *k*=2, (a) T_k (9x2), (b) S_k (2x2) e (c) D_k (8x2)

Através dessas matrizes, relacionamentos latentes são estabelecidos, podendo ser do tipo termo–termo, documento–documento ou termo–documento. A opção termo–termo será discutida na seção sobre modelos de correlação. Segundo Ding (2000), o produto dos pontos entre dois vetores é geralmente admitido como uma medida de correlação aceitável, definido como:

$$sim(x_1, x_2) = x_1 \cdot x_2 , \tag{4}$$

Nesse sentido, a similaridade entre dois vetores que representam documentos pode ser obtida através de $X^T X$. $X^T X$ é, portanto, a matriz de similaridades entre todos os pares de documentos. Aplicações de agrupamentos de documentos estão entre as possibilidades de utilização. A matriz *X,* resultado da Equação 3, é apresentada na Tabela 6.

0,4837	0,5154	0,3556	0,0000..	1,0761	0,5351	0,5027	0,4542
0,2292	0,4440	0,1953	-0,0000..	0,5835	0,4969	0,1429	0,2163
0,6220	0,0549	0,3758	0,0000..	1,1602	-0,0521	0,9365	0,5810
-0,0033	0,7742	0,1018	-0,0000..	0,2787	0,9433	-0,3745	0,0009
0,2293	0,9199	0,2591	-0,0000..	0,7586	1,0763	-0,0840	0,2187
0,3046	0,9723	0,3107	-0,0000..	0,9159	1,1257	0,0075	0,2893
0,2200	0,5700	0,2067	-0,0000..	0,6129	0,6520	0,0685	0,2083
0,4217	-0,0068	0,2488	0,0000..	0,7703	-0,0889	0,6559	0,3936
0,5385	0,0037	0,3194	0,0000..	0,9882	-0,0985	0,8316	0,5027

Tabela 6 - Matriz *X* (9x8) resultante de $X = TSD^{'} \cong T_k S_k D_k^T$

Outra possibilidade é a aplicação de SVD na área de recuperação de informação, em que, através de um conjunto de termos informados pelo usuário, os documentos mais

relevantes são recuperados. Como exemplo, considere o vetor q={0,1,1,0,0,0,0,0,0} (normalizado para tamanho 1), em que cada posição representa um termo na matriz termo–documento (neste caso, conhecimento e mineração). O valor 1 significa a existência do termo no vetor de consulta, do contrário, 0. A equação utilizada na normalização do vetor de consulta baseada nos k fatores mais relevantes é definida como:

$$D_q = X_q^{'} T_k S_k^{-1}, \text{ ou } dq = q^T \cdot T_k \cdot inv(S_k) \quad (5)$$

Aplicando-se a Equação 5 sobre o vetor de consulta, o resultado será D_q= {0.1651, -0,1784}. O próximo passo calcula a distância ou a similaridade entre cada linha da matriz D_k e D_q, utilizando a equação do co-seno (Equação 1). Como exemplo, o vetor da consulta (D_q) e o vetor do documento $D1$={0,2762, -0,2208} são considerados, e os resultados parciais para o cálculo, apresentados na Tabela 7.

T	t^2	q	q^2	t*q
0,2762	0,0763	0,1651	0,0273	0,0456
-0,2208	0,0488	-0,1784	0,0318	0,0394
	0,1250		**0,0591**	**0,0850**

Tabela 7 - Demonstração parcial do cálculo para a equação do co-seno, $\sigma(dq, D)$

Utilizando-se a tabela acima, o resultado para o cálculo do co-seno entre o vetor de consulta (D_q) e o documento $D1$ será:

$$\cos\theta = \frac{0,0850}{\sqrt{0,1250} \times \sqrt{0,0591}} = 0,9888$$

A Tabela 8 apresenta as similaridades (ordenadas) entre o vetor de consulta e a matriz de documentos, sendo $D1$, $D8$ e $D7$ os mais similares em relação ao vetor de consulta.

Documento	$\sigma(Dq, D_i)$
*D*1	0,9888
*D*8	0,9874
*D*7	0,9442
*D*5	0,8921
*D*3	0,8748
*D*4	0,7167
*D*2	-0,1050
*D*6	-0,1978

Tabela 8 - Similaridade entre o vetor de consulta e a matriz de documentos (D_k)

2.4 MODELOS BASEADOS EM CO-OCORRÊNCIA

Um dos problemas identificados em sistemas tradicionais de recuperação e análise de informação textual reside na falta de contexto, no qual elementos textuais que ocorram ao longo da coleção de documentos não possuem relacionamentos entre si. Tal fato pode obscurecer o significado latente nesse tipo de estrutura. Modelos baseados em co-ocorrência podem promover a resposta para se atingirem melhoramentos tanto na representação de documentos quanto no mapeamento de conhecimento implícito em bases textuais. Entre os modelos, citam-se os advindos da estatística descritiva, tais como o teste *t*, o *Chi-square* x^2 e o *Z score* (MANNING; SCHÜTZE, 1999), e os modelos advindos da teoria da informação, tais como Informação Mútua (MI) (CHURCH; HANKS, 1990; CHURCH; GALE, 1991) e *Phi-squared* (ϕ^2) (CHURCH; GALE, 1991), ambos com o intuito de identificar seqüências naturais de palavras (*collocations*). Cita-se ainda o modelo LSI, que objetiva capturar a estrutura semântica de coleções de documentos através da correlação de termos e documentos (DEERWESTER et al., 1990; DING, 2000).

Tais modelos partem do pressuposto de que é possível estabelecer estatisticamente uma possível relação entre palavras, analisando-se suas freqüências conjuntas, ou seja, as co-ocorrências. Além das co-ocorrências, destaca-se ainda a distância entre duas palavras. Segundo Croft et al. (1991), a distância entre palavras é uma forte evidência para a

presença de relacionamentos frasais. Para tal, utilizam-se janelas indicando o número de palavras à direita e à esquerda de uma determinada característica. Tal conceito pode ser facilmente generalizado para qualquer elemento textual, como, por exemplo, entidades. A seguir serão detalhados os modelos de correlação estudados assim como os principais conceitos utilizados por esses modelos.

2.4.1 FREQÜÊNCIA

Uma das maneiras mais simples de se estabelecer a relação entre dois elementos textuais, ainda que imprecisa, é a freqüência conjunta. O fato de duas palavras, ou qualquer outro elemento textual, aparecerem freqüentemente em uma determinada coleção de documentos demonstra a evidência de relacionamento.

Entretanto, selecionar simplesmente bigramas freqüentes em uma coleção não é uma solução interessante. Veja o exemplo a seguir apresentado na Tabela 9.

$C(t1,t2)$	$t1$	$t2$
80874	of	the
58841	in	the
26430	to	the
....		
12622	from	the
11428	New	York

Tabela 9 - Exemplo de freqüências conjuntas extraídas de uma coleção de documentos
FONTE: Extraído de: JUSTESON; KATZ, 1995.

Uma alternativa simples seria a eliminação de bigramas formados simplesmente por palavras constantes em uma tabela de controle (*stop lists*). Outra alternativa que tende a melhorar esses resultados é proposta por Justeson e Katz (1995), em que são utilizados padrões que identificam prováveis estruturas frasais. Basicamente existem três unidades que compõem os padrões, sendo A para um adjetivo, N para um nome e P para uma preposição. Utilizando-se esses padrões, os resultados são incrementados

consideravelmente (Tabela 10), sendo "*New York*" agora classificada com maior relevância.

$C(t1,t2)$	$T1$	$t2$	Padrão
11428	New	York	AN
5412	Los	Angeles	NN
3301	last	year	AN
....			

Tabela 10 - Exemplo de freqüências conjuntas aplicando-se o filtro proposto por Justeson e Katz
FONTE: Extraído de: JUSTESON; KATZ, 1995.

2.4.2 MÉDIA E VARIÂNCIA

Embora o uso de freqüência conjunta releve indícios para a formação de estruturas frasais, muitas dessas estruturas ocorrem de maneira mais flexível, em que palavras são conectadas através de janelas. A quantidade de palavras que aparece entre outras duas palavras varia, e a distância entre elas não é a mesma. A utilização de janelas (quantidade de palavras em cada um dos lados de uma determinada palavra) oferece a solução.

Como exemplo consideram-se duas palavras $t1$ e $t2$ que ocorrem com diferentes deslocamentos ao longo da coleção de documentos, sendo esses deslocamentos 5, 5, 3, 4, 4, respectivamente. Nesse sentido, a média e a variância podem determinar o grau de relacionamento entre as palavras. A média é computada utilizando-se os deslocamentos, como mostrado a seguir.

$$\frac{1}{5}(5+5+3+4+4) = 4.2$$

A variância informa o grau de desvio dos deslocamentos a partir da média, sendo estimada conforme a seguinte equação:

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} \qquad (6)$$

onde n é o número de vezes que as duas palavras co-ocorrem, d_i é o deslocamento da ith co-ocorrência, e $\bar{d}$ é a média dos deslocamentos. Caso os deslocamentos sejam sempre os mesmos, a variância será zero. Do contrário, se os deslocamentos acontecem

aleatoriamente, ou seja, não configuram um padrão de relacionamento, a variância será alta.

Nesse sentido o desvio padrão $s = \sqrt{s^2}$, a raiz quadrada da variância, é utilizado para avaliar a variabilidade dos deslocamentos entre duas palavras ou qualquer outra estrutura textual. Seguindo-se o exemplo acima, o resultado seria:

$$s = \sqrt{\frac{1}{4}((5-4.2)^2 + (5-4.2)^2 + (3-4.2)^2 + (4-4.2)^2 + (4-4.2)^2)} \approx 0.8366$$

A informação provida pela média e pela variância, ou seja, a distribuição da distância entre duas palavras na coleção de documentos, pode ser utilizada na determinação de estruturas frasais com baixo desvio padrão. Valores de desvios baixos indicam que duas palavras tendem a ocorrer quase sempre na mesma distância, enquanto que o valor zero indica que duas palavras ocorrem exatamente sempre na mesma distância. Por sua vez, valores de desvios altos indicam relacionamentos entre pares pouco relevantes.

2.4.3 TESTE DE HIPÓTESE

Embora altas freqüências e baixas variâncias possam indicar a constituição de estruturas mais complexas, não existe garantia de que isso conduza a resultados melhores dos que aqueles obtidos ao acaso. O que se deseja realmente é identificar padrões que ocorrem mais freqüentemente do que o acaso. Avaliar se algo é ou não um evento ao acaso é um dos problemas clássicos da estatística (MANNING; SCHÜTZE, 1999), identificado como teste de hipóteses.

A hipótese nula H_0 indica inicialmente que não existe associação entre duas palavras *A* e *B*. Contudo, se a probabilidade *p* de um evento é inferior a determinado valor

crítico ou nível de significância[12] (p < 0.05, 0.01, 0.005, ou 0.001, sendo os valores críticos iguais a 1.645, 2.326, 2.576 e 3.091, respectivamente), pode-se rejeitar H_0, ou seja, existe uma associação ou um relacionamento entre A e B, do contrário, se a probabilidade p é superior a determinado valor crítico, H_0 não pode ser rejeitado.

2.4.4 TESTE *T*

Este teste tem sido utilizado extensivamente na identificação de "*collocations*". O teste t informa o quão provável ou improvável é a ocorrência de determinado evento. Através da média e da variância, a hipótese nula é analisada informando que a amostra é composta a partir de uma distribuição com média μ. O resultado é, portanto, obtido através da análise das diferenças entre as médias observadas e esperadas, normalizadas pela variância dos dados. Desse modo, a probabilidade da amostra para a estatística t é computada como:

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \tag{7}$$

onde $\overline{x}$ é a média da amostra, s^2 é a variância da amostra, N é o tamanho da base (quantidade de pares de palavras (bigramas) existentes na coleção de documentos) e μ é a média da distribuição. Se o teste t é grande o suficiente, a hipótese nula pode ser rejeitada, ou seja, a determinação da relação entre elementos textuais pode ser confirmada.

Tradicionalmente este teste é aplicado à amostra de dados. Entretanto, no contexto da identificação de "*collocations*", existe uma forma padronizada para estender o teste t para o uso de proporções e contagens. Assim, uma coleção de documentos é tratada como uma longa seqüência de N pares de palavras (bigramas). As amostras são obtidas aplicando-se 1 quando o bigrama de interesse ocorre, caso contrário, 0.

[12] O nível de significância de 0.05 é geralmente aceito nas ciências experimentais como a evidência para se validar ou rejeitar a hipótese nula.

Estimando-se a máxima probabilidade, torna-se possível o cálculo das probabilidades de cada componente do bigrama. Para exemplificar, as palavras *t*1="inteligência" e *t*2="artificial" são levadas em consideração. Na coleção de documentos *t*1 ocorre 14.902 vezes, *t*2 ocorre 6.484 vezes, em um total de 15.806.525 bigramas.

$$P(t1) = \frac{14.902}{15.806.252}$$

$$P(t2) = \frac{6.484}{15.806.252}$$

A hipótese nula informa inicialmente que as ocorrências de *t*1 e *t*2 são independentes.

$$\begin{aligned} H_0 : P(t1t2) &= P(t1)P(t2) \\ &= \frac{14.902}{15.806.252} \times \frac{6.484}{15.806.252} \approx 3.8675 \times 10^{-7} \end{aligned}$$

Como exemplo assume-se que existam 32 ocorrências de "inteligência artificial" entre os 15.806.252 pares da coleção de documentos. Assim, para o exemplo, a média seria: $\overline{x} = \frac{32}{15.806.252} \approx 2.02452 \times 10^{-6}$. Aplicando-se esses resultados na Equação 7, o valor do teste *t* seria:

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{P(t1t2) - P(t1)P(t2)}{\sqrt{\frac{P(t1t2)}{N}}} \approx \frac{2.02452 \times 10^{-6} - 3.8675 \times 10^{-7}}{\sqrt{\frac{2.02452 \times 10^{-6}}{15.806.252}}} = 4.576208$$

O valor do teste *t*, considerando-se o valor crítico para o nível de probabilidade de $\alpha = 0.005$, é maior do que 2.576. Desse modo, a hipótese nula pode ser rejeitada, ou seja, "inteligência" e "artificial" ocorrem independentemente. Esse valor indica que "inteligência artificial" não é meramente composta ao acaso e que existe significado para que esses termos sejam utilizados conjuntamente, ou seja, que formem "*collocations"*.

2.4.5 *TESTE DE PEARSON - CHI-SQUARE (χ^2)*

Ao contrário do teste *t,* que assume que as probabilidades são normalmente distribuídas (CHURCH; MERCER, 1993), o teste χ^2 não possui essa dependência. Basicamente χ^2 é uma técnica estatística utilizada para determinar se a distribuição das freqüências observadas difere das freqüências esperadas. Se a diferença entre as freqüências observadas e esperadas é alta, então a hipótese nula de independência pode ser rejeitada. Sua aplicação baseia-se na utilização de uma tabela 2*2 (tabela de contingência), como a apresentada a seguir.

	w_2	$\overline{w}_2$
w_1	a	b
$\overline{w}_1$	c	d

Tabela 11 - Tabela de contingência de 2x2

onde a célula *a* indica o número de vezes que w_1 e w_2 ocorrem conjuntamente, *b* indica o número de vezes que w_1 ocorre mas w_2 não, *c* é o número de vezes que w_2 ocorre mas w_1 não, e, finalmente, *d* é o tamanho da coleção de documentos menos o número de vezes que nem w_1 e nem w_2 ocorrem, sendo $d=N-a-b-c$, onde N é o tamanho da base.

A estatística χ^2 soma a diferença entre os valores observados e esperados divididos pelos valores esperados, como definido a seguir.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = \frac{\left(a - \frac{(a+b)(a+c)}{N}\right)^2}{\frac{(a+b)(a+c)}{N}} + \frac{\left(b - \frac{(a+b)(b+d)}{N}\right)^2}{\frac{(a+b)(b+d)}{N}} + \frac{\left(c - \frac{(c+d)(a+c)}{N}\right)^2}{\frac{(c+d)(a+c)}{N}} + \frac{\left(d - \frac{(c+d)(b+d)}{N}\right)^2}{\frac{(c+d)(b+d)}{N}} = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad (8)$$

Seja a Tabela 12 a representação da distribuição para as palavras “inteligência” e “artificial”:

	w_2 =artificial	$\overline{w}_2 \neq$ artificial
w_1 = inteligência	**32**	14.902-32=**14.870**
$\overline{w}_1 \neq$ inteligência	6.484-32=**6.452**	15.806.252-14.870-6.452-32=**15.784.898**

Tabela 12 - Exemplo de tabela de contingência para a dependência das palavras *t*1="inteligência" e *t*2="artificial"[13]

Aplicando-se a Equação 8 sobre os valores demonstrados na Tabela 12, teria-se:

$$\chi^2 = \frac{15.806.252 \times (32 \times 15.784.898 - 14.870 \times 6.452)^2}{(32+14.870) \times (32+6.452) \times (14.870+15.784.898) \times (6.452+15.784.898)} = 109.77$$

A hipótese nula indica que as ocorrências das palavras *t*1="inteligência" e *t*2="artificial" H_0: *P*(*inteligência artificial*) são independentes. Analisando-se a distribuição de χ^2, pode-se verificar que para o nível de probabilidade de $\alpha = 0.05$ o valor crítico é $\chi^2 = 3.841$ para um grau de liberdade (${}_{(0.05)}\chi^2{}_{(1)} = 3.841$). Se χ^2 está abaixo de ${}_{(0.05)}\chi^2{}_{(1)} = 3.841$, H_0 não pode ser rejeitada. Se χ^2 está acima de ${}_{(0.05)}\chi^2{}_{(1)} = 3.841$, H_0 pode ser rejeitada, ou seja, existe um relacionamento entre *t*1 e *t*2.

2.4.6 *PHI-SQUARED* (ϕ^2)

Segundo Conrad e Utt (1994), ϕ^2 tende a favorecer associações com alta freqüência. Similar ao método anterior, uma tabela de contingência é utilizada. *Phi-squared* (CHURCH; GALE, 1991) é definida como:

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{9}$$

onde $0 \leq \phi^2 \leq 1$.

Visto que tanto *Chi-square* quanto *Phi-squared* podem considerar associações que estejam dentro de uma determinada janela de palavras, a célula *d* da tabela de

[13] Existem 32 ocorrências para "inteligência artificial" na coleção de documentos, 14.870 bigramas, onde a primeira palavra é "inteligência" mas a segunda não é "artificial", 6.452 bigramas, em que a segunda palavra é "artificial" mas a primeira não é "inteligência", e 15.784.898 bigramas que não contém nenhuma das palavras.

contingência pode ser calculada de maneira diferente, sendo $d = \frac{N}{w} - a - b - c,$ onde w é o tamanho da janela de texto (o conceito de janela será discutido na seção seguinte).

2.4.7 *INFORMAÇÃO MÚTUA E DERIVAÇÃO*

A Informação Mútua (MI) possui motivação na teoria da informação e tem sido aplicada na identificação do nível de associação entre palavras utilizando as informações de co-ocorrências na coleção de documentos (CHURCH; HANKS, 1990). Informação Mútua compara a probabilidade de um par de palavras, ou qualquer outra unidade lingüística, aparecer mais freqüentemente de maneira conjunta do que apareceria isoladamente. Essa medida cresce à proporção que a freqüência conjunta também cresce. Se um determinado termo tende a ocorrer individualmente, então MI será um número negativo.

A fórmula padronizada para o cálculo de MI é definida como:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N} \times \frac{f(y)}{N}} \tag{10}$$

onde *P(x,y)* é a probabilidade de duas palavras *x* e *y* ocorrerem conjuntamente, *P(x)* e *P(y)* são as probabilidades de *x* e *y* ocorrerem individualmente, e *N,* o tamanho da base. Quando existe um relacionamento forte entre as palavras (características), *I(x,y)* será maior que 0.

Para exemplificar o cálculo, a máxima probabilidade é utilizada na determinação da probabilidade de dois eventos que ocorrem conjuntamente. Considere o seguinte exemplo:

$$I(Inteligência, Artificial) = \log_2 \frac{\frac{32}{15.806.252}}{\frac{14.902}{15.806.252} \times \frac{6.484}{15.806.252}} \approx 2.38$$

Mais especificamente, MI informa que a quantidade de informação da ocorrência de “Inteligência” na posição *i* da coleção aumenta em 2.38 *bits* se é aceito que “Artificial” ocorre na posição *i* + 1, ou vice-versa. O problema encontrado com MI, assim como em outros métodos derivados, refere-se ao tratamento de eventos de baixa freqüência.

Melhoramentos sobre MI têm sido desenvolvidos, entre eles, citam-se os trabalhos de Vechtomova et al. (2003) e Wang e Vechtomova (2005), que introduzem um parâmetro adicional: o tamanho da janela entre o par de palavras. Uma janela é definida como um número fixo de palavras à direita e à esquerda de uma palavra ***x***, ou seja, cada janela é estabelecida ao redor de uma determinada palavra (Figura 2). Em um modelo ideal, ambos os lados da janela possuem o mesmo tamanho. Todavia isso nem sempre acontece. Como regra geral, a relação será estabelecida considerando-se as *S* palavras à esquerda e as *S* palavras à direita da palavra fixada.

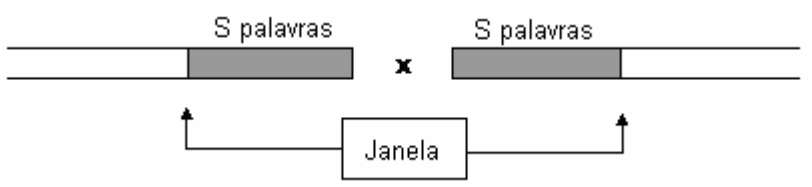


Figura 2 - Janela ao redor da palavra ***x***, definida como *S* palavras à esquerda e à direita de ***x***
FONTE: Adaptado de: VECHTOMOVA et al., 2003.

Contudo, por duas razões, as janelas utilizadas são freqüentemente menores do que as sugeridas por essa distância. A janela pode ser truncada se: (a) alcança o limite do documento (Figura 3); ou (b) alcança uma outra ocorrência da palavra ***x*** (Figura 4 e Figura 5). Nesse caso, se outra instância da palavra é encontrada antes da palavra ***x*** ou após ela, a janela é truncada nesse ponto para evitar sobreposições.

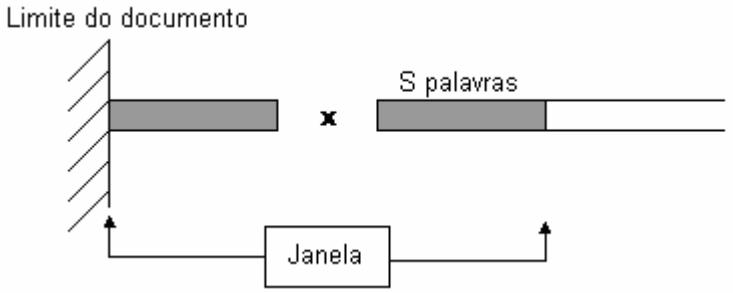


Figura 3 - Janela truncada por ter atingido o limite inferior do documento

FONTE: Adaptado de: VECHTOMOVA et al., 2003.

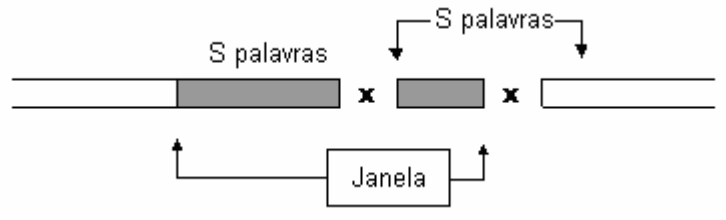


Figura 4 - Janela truncada à direita por ter alcançado outra ocorrência da palavra *x* após o ponto central

FONTE: Adaptado de: VECHTOMOVA et al., 2003.

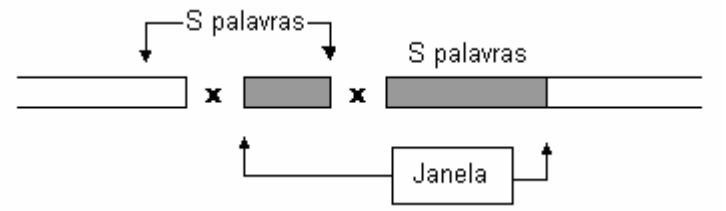


Figura 5 - Janela truncada à esquerda por ter alcançado outra ocorrência da palavra *x* antes do ponto central

FONTE: Adaptado de: VECHTOMOVA et al., 2003.

A principal diferença dessa abordagem é a sua assimetria. A medida de MI tradicional é simétrica, isto é $I(x,y) = I(y,x)$, assim como probabilidades conjuntas são simétricas, $P(x,y) = P(y,x)$. Para se estimar a probabilidade de ocorrência de y em uma janela ao redor de x, a média das janelas é calculada através da equação:

$$v_x = \frac{\sum_{i=1}^{f(x)} W_i}{f(x)} \tag{11}$$

onde W_i é a janela observada ao redor da ith instância de $\boldsymbol{x}$ na coleção de documentos, e $f(x)$, a freqüência de x na coleção de documentos.

Sendo assim, a equação modificada de MI é definida como:

$$I_v(x,y) = \log_2 \frac{P_v(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{f(x,y)}{Nv_x}}{\frac{f(x)f(y)}{N^2}} \quad (12)$$

onde $f(x,y)$ é a freqüência do conjunto de x e y na coleção de documentos, $f(x)$ e $f(y)$ são as freqüências independentes de x e y, v_x é o tamanho médio da janela ao redor de x, e N é o tamanho da base.

2.4.8 *Z SCORE*

O *Z score* promove uma indicação sobre a validade da associação entre elementos textuais medindo-se a distância dos desvios padrão entre as freqüências observadas das ocorrências de y em torno de x e as freqüências esperadas. Ao contrário da medida de Informação Mútua, que sob a condição de possuir baixas freqüências pode conduzir a valores elevados, o *Z score* não produz esse efeito, pois a variância das probabilidades será elevada (VECHTOMOVA et al., 2003). A equação *Z score* pode ser definida como:

$$Z(x,y) = \frac{O-E}{\sqrt{E}} = \frac{f(x,y) - \frac{f(x)f(y)}{N}}{\sqrt{\frac{f(x)f(y)}{N}}} \quad (13)$$

Vechtomova et al. (2003) propõem uma derivação levando em conta uma determinada janela em torno de x. A hipótese nula, segundo os autores, indica que a presença de x não prediz a presença ou a ausência de y na janela, uma vez que a probabilidade de y ocorrer entre a janela e outra localização do documento é a mesma.

De maneira geral, o número de localizações que podem conter o termo y associado com x é $v_x f(x)$. A probabilidade de qualquer uma dessas localizações possuir y é determinada por $f(y)/N$. Assim, o número esperado de ocorrência de y é a média da distribuição binomial[14], $v_x f(x)\, f(y)/N$. *Z* score é calculado como:

[14] Uma distribuição binomial representa uma distribuição de probabilidade do número de sucessos obtidos em uma seqüência de n independentes experimentos, na qual cada um produz sucessos com probabilidade p.

$$Z(x,y) = \frac{O-E}{\sqrt{E}} = \frac{f(x,y) - \frac{v_x f(x) f(y)}{N}}{\sqrt{\frac{v_x f(x) f(y)}{N}}} \quad (14)$$

onde *f(x,y)* é a freqüência de *x* e *y* na coleção de documentos, *f(x)* e *f(y)* são as freqüências das ocorrências independentes de *x* e *y,* respectivamente, v_x é o tamanho médio da janela ao redor de *x* na coleção de documentos, e *N* é o tamanho da base.

2.4.9 INDEXAÇÃO SEMÂNTICA LATENTE

Como discutido na Seção 2.3.2, LSI pode ser aplicada na identificação de correlações, seja entre documento–documento ou termo–documento. Adicionalmente, a correlação pode ser estendida para termo–termo. Similarmente ao modelo documento–documento, o produto dos pontos entre dois vetores representando termos é definido como:

$$sim(t^1, t^2) = t^1 \cdot t^2, \quad (15)$$

A equação calcula as co-ocorrências entre termos considerando todos os documentos constantes na coleção. Através do resultado da Equação 3 (matriz *X* apresentada na Tabela 6), o produto dos pontos XX^T (Tabela 13) pode ser aplicado para se produzir a matriz termo–termo.

2,5294	1,4730	2,4181	1,0505	2,1266	2,4822	1,6112	1,5788	2,0330
1,4730	0,9424	1,1509	0,9409	1,5244	1,7197	1,0803	0,7264	0,9426
2,4181	1,1509	3,0945	0,0028	1,1630	1,5388	1,1081	2,0967	2,6777
1,0505	0,9409	0,0028	1,7175	1,9962	2,0980	1,2221	-0,0957	-0,0948
2,1266	1,5244	1,1630	1,9962	2,7549	3,0140	1,8350	0,6746	0,8934
2,4822	1,7197	1,5388	2,0980	3,0140	3,3247	2,0417	0,9235	1,2128
1,6112	1,0803	1,1081	1,2221	1,8350	2,0417	1,2649	0,6814	0,8897
1,5788	0,7264	2,0967	-0,0957	0,6746	0,9235	0,6814	1,4262	1,8198
2,0330	0,9426	2,6777	-0,0948	0,8934	1,2128	0,8897	1,8198	2,3225

Tabela 13 - Produto dos pontos XX^T representando as dimensões de termos

Para se calcular a similaridade entre as palavras, a equação do co-seno pode ser utilizada (Equação 1). Seja a palavra de origem *A*=Carvão, deseja-se saber qual a correlação das demais palavras pertencentes à Tabela 1. Como exemplo, a dimensão 9 (Carvão) e a dimensão 1 (Descoberta) são consideradas, e os resultados parciais para o cálculo, apresentados na Tabela 14.

T	t^2	q	q^2	t*q
2,5294	6,3979	2,0330	4,1331	5,1423
1,4730	2,1697	0,9426	0,8885	1,3884
2,4181	5,8472	2,6777	7,1701	6,4749
1,0505	1,1036	-0,0948	0,0090	-0,0996
2,1266	4,5224	0,8934	0,7982	1,8999
2,4822	6,1613	1,2128	1,4709	3,0104
1,6112	2,5960	0,8897	0,7916	1,4335
1,5788	2,4926	1,8198	3,3117	2,8731
2,0330	4,1331	2,3225	5,3940	4,7216
	35,4238		**23,9669**	**26,8446**

Tabela 14 - Demonstração parcial do cálculo para a equação do co-seno

Utilizando-se a tabela acima, o resultado para o cálculo do co-seno para as palavras "Carvão" e "Descoberta" seria:

$$\cos\theta = \frac{26,8446}{\sqrt{35,4238} \times \sqrt{23,9669}} = 0,9213$$

O cálculo é repetido para as demais palavras (veja Tabela 1) objetivando determinar as suas correlações. Sendo assim, a palavra "Carvão" terá as palavras "Turbina", "Mineração" e "Descoberta" como as mais similares/correlacionadas (Tabela 15).

Índice	Palavra	sim(t^1,t^2)
9	Carvão	1,0000
8	Turbina	0,9999
3	Mineração	0,9992
1	Descoberta	0,9213
2	Conhecimento	0,8340
7	Radiação	0,7803
6	Retroalimentação	0,7402
5	Usuário	0,6968
4	Interface	0,4183

Tabela 15 - Similaridades entre o termo "Carvão" e os demais termos da coleção de documentos

2.5 DESCOBERTA DE CONHECIMENTO

Os avanços da era da informação têm criado novos desafios no desenvolvimento de técnicas e métodos voltados à extração de padrões relevantes na crescente massa de dados tanto no nível organizacional quanto na *Internet*. Igualmente, esses avanços têm promovido meios de acessar a informação de maneira rápida e eficiente. Embora isso seja de grande valia para as organizações, tornam-se necessários mecanismos que promovam suporte ao entendimento das bases de dados textuais. Desse modo, a habilidade de extrair conhecimento útil a partir de bases de dados voltado ao melhoramento do processo organizacional torna-se estratégica. Por "conhecimento útil" entende-se aquele que pode ser acionável, ou seja, o conhecimento que gera ações.

A área de descoberta de conhecimento baseia-se em duas tendências: uma avalanche de informação e um questionamento sobre essa informação (HAIR et al., 1998). Nesse sentido, a análise de dados passa a ter um caráter mais exploratório, visando identificar ou explicitar conhecimento latente em bases de dados. Essa tarefa é de responsabilidade da área de Descoberta de Conhecimento em Bases de Dados (KDD). KDD (Figura 6) constitui-se no processo envolvendo a seleção, o pré-processamento, a transformação do dado, a utilização de algoritmos especializados e a geração de conhecimento (FAYYAD et al., 1996a). O modelo possui processos recorrentes entre as fases, isto é, a cada avaliação da fase corrente, a(s) fase(s) anterior(es) pode(m) sofrer ajuste(s). Segundo Mack e Hehenberger (2002), os métodos de descoberta de conhecimento evolvem a compilação e a integração de diferentes formatos de dados para criar um contexto interpretativo com a intenção de promover suporte ao entendimento e implicações do conhecimento não relevado em bases de dados.

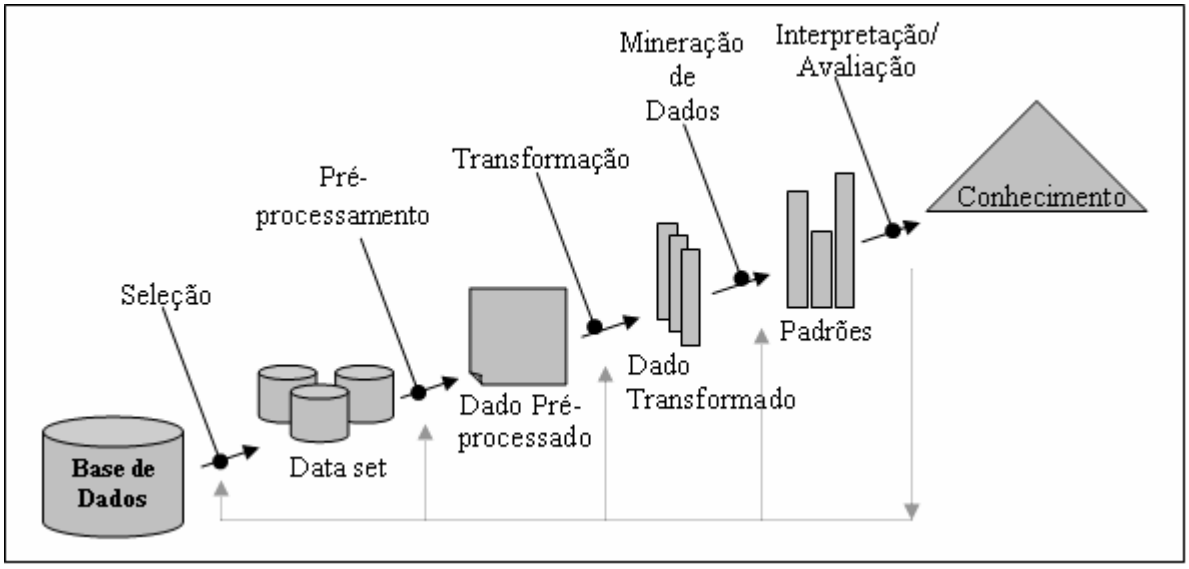


Figura 6 - Processo de KDD

FONTE: Adaptado de: FAYYAD et al., 1996a.

Uma das fases mais importante do processo é a Mineração de Dados (DM), sendo responsável pela aplicação de algoritmos com o propósito de identificar padrões em uma base de dados (FAYYAD, 1996b). A Mineração de Dados promove uma evolução na perspectiva tradicional de análise de dados, permitindo descobrir novos padrões e/ou validar aqueles muitas vezes identificados de maneira *ad hoc*. Essas análises são geralmente executadas sobre uma grande quantidade de dados pré-processados e armazenados por *Data Warehouses*/*Data Marts*. De acordo com Harrison (1998), DM constitui-se na exploração e na análise, através de meios automáticos ou semi-automáticos, de grandes quantidades de dados com o intuito de descobrir padrões relevantes.

Possui como metas primárias a (a) previsão, através das quais variáveis (atributos ou dimensões em bases de dados) são utilizadas para determinar, por meio de análises históricas, o que deve acontecer em um momento futuro e (b) a descrição, através da busca por padrões, de modo a gerar mapas de conhecimento (BERRY; LINOFF, 1997). Levando-se em conta tais metas, pode-se observar a real importância desse processo dentro de uma organização, sendo que essas metas devem prever algumas fases. Harrison (1998) identifica quatro fases no processo de DM, entre elas: (a) identificação de problemas e áreas para as quais a análise de

dados possa agregar valor; (b) transformação dos dados em informações acionáveis usando técnicas de DM; (c) ação sobre a informação e, a partir dela, implementação de melhorias nos processos que regem o relacionamento da empresa com os seus colaboradores, consumidores e fornecedores; e (d) aferição dos resultados obtidos através da aplicação das técnicas de DM. Esta fase proporciona o *feedback* para o aumento constante na qualidade dos resultados.

Apesar de sua utilidade e aplicabilidade, a DM promove ferramentas que atendem parcialmente à demanda pela descoberta de conhecimento. A rápida difusão de bases textuais e a necessidade de se recuperarem informações úteis requerem abordagens diferenciadas para se lidar com o problema. Estudos indicam que grande quantidade de informação dentro das organizações encontra-se na forma de textos (TAN, 1999; DÖRRE et al., 1999). Nesse sentido, a análise quantitativa tem se movido para uma nova perspectiva, a Descoberta de Conhecimento a partir de Bases Textuais (KDT) e a Mineração de Textos (TM).

Similar ao KDD, KDT refere-se ao processo de maneira geral, enquanto que TM representa o processo de extração de padrões relevantes e não triviais a partir de bases de dados semi ou não estruturadas. Pode-se assim afirmar que a TM é uma extensão da Mineração de Dados tradicional. Conceitos similares podem ser encontrados em Wohl (1998), Dörre et al. (1999) e Witten et al. (1999). De acordo com Nasukawa e Nagano (2001), TM é a versão textual da DM, compondo-se de técnicas de NLP para extrair conceitos de textos, análises estatísticas para recuperar padrões e técnicas de visualização para permitir análises interativas (Figura 7).

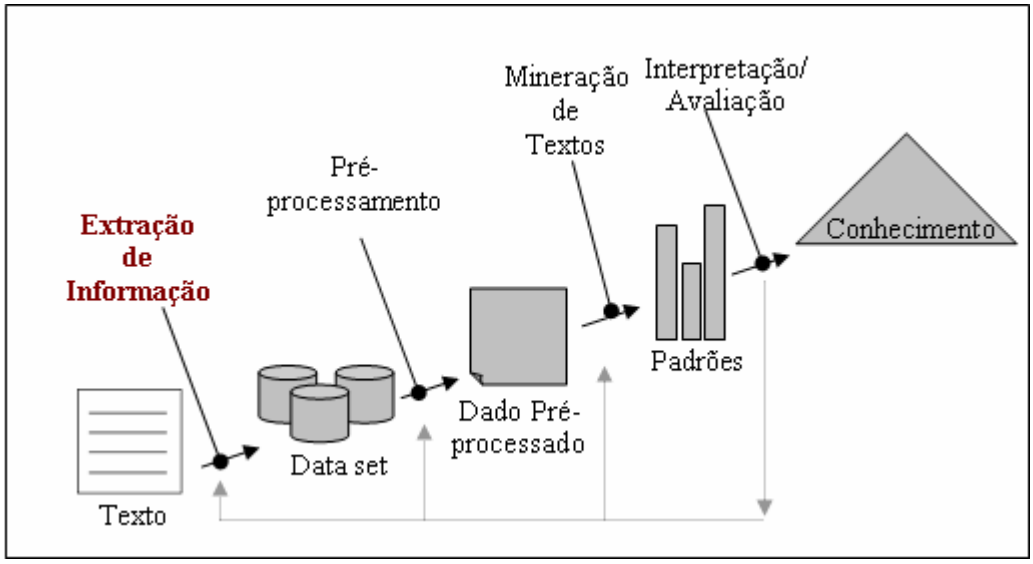


Figura 7 - Processo de KDT

FONTE: Adaptado de: MOONEY; NAHM, 2005.

Entretanto, tal afirmação é incompleta por não prever a dificuldade na manipulação de dados de alta dimensionalidade. Documentos textuais podem gerar representações com diferentes tamanhos e atingem centenas e mesmo milhares de dimensões. Nesse sentido a habilidade de se lidar com estruturas esparsas e mesmo a capacidade de reduzir a dimensionalidade desses vetores tornam-se de grande importância para que sejam obtidos resultados satisfatórios.

Embora essa abordagem promova ferramentas para a análise do crescente volume de informações textuais, através da descoberta de padrões relevantes, a maioria dos algoritmos é baseada simplesmente em termos, e não necessariamente em estruturas mais complexas que possibilitem suporte a aplicações de engenharia e gestão do conhecimento. A evolução desse modelo envolve o uso de ontologias e modelos de co-ocorrências de modo a determinar padrões e relacionamentos entre elementos textuais. Como exemplos citam-se o trabalho de Bi et al. (2003), que propõem a integração de algoritmos de regras de associação e ontologias. Uma abordagem mais robusta baseada na extração de conceitos e utilização desses conceitos na identificação de padrões e regras tem sido proposta por Feldman et al. (1998) e Loh e Wives (2000).

Em outro estudo mais diretamente aplicado à área de recuperação de informação, Andreasen et al. (2003) propõem um modelo para explorar conhecimento representado na forma de ontologia em resposta a consultas em bases de informação. Cita-se ainda o trabalho de Gonçalves et al. (2005), que se utilizam de ontologias para expandir o espaço vetorial[15] com o objetivo de melhorar o contexto de representação de cada documento na base de dados. Nesse sentido, TM é também identificada como um conjunto adicional de funções que podem incrementar sistemas de recuperação de informação.

Entre essas funções/tarefas, citam-se agrupamento, classificação, sumarização de documentos, descobrimentos de regras e análise de ligações. A Tabela 16 demonstra as principais funções/tarefas, os algoritmos/técnicas e os exemplos de aplicações disponíveis na DM tradicional, que com as devidas adaptações são também aplicáveis à TM. A escolha de uma ou de outra depende essencialmente do negócio, da aplicação e da quantidade e qualidade dos dados disponíveis.

Funções/Tarefas	Algoritmos/Técnicas	Aplicações
Associação	Estatística, teoria dos conjuntos	Análise de mercados
Classificação	Árvores de decisão, redes neurais, algoritmos genéticos	Controle de qualidade, avaliação de riscos
Agrupamentos	Redes neurais, estatística	Segmentação de mercado
Modelagem	Regressão linear e não-linear, redes neurais	*Ranking* de clientes, controle de processos, modelos de preços
Previsão de séries temporais	Estatística, redes neurais	Previsão de vendas, controle de inventário
Padrões seqüenciais	Estatística, teoria dos conjuntos	Análise de mercado sobre o tempo

Tabela 16 - Funções/tarefas da Mineração de Dados

FONTE: BIGUS, 1996.

De acordo com Chen et al. (1996), diferentes esquemas de classificação têm sido utilizados para categorizar métodos de mineração de dados quanto ao tipo de base de

[15] A expansão vetorial ocorre pela localização, em uma determinada ontologia, de sinônimos, generalizações e especializações de determinado termo pertencente ao vetor que representa o documento. Esses novos termos localizados são então adicionados ao contexto original do documento.

dados a ser estudada, ao tipo de conhecimento a ser descoberto e ao tipo de técnica a ser utilizada, como descrito a seguir:

- *Tipo de base de dados*: os sistemas de mineração de dados podem ser classificados segundo o tipo da base de dados em que estão sendo executados, ou seja, bases de dados relacionais, dimensionais, textuais.
- *Tipo de conhecimento a ser extraído*: diferentes tipos de conhecimento podem ser identificados através de sistemas de mineração de dados, incluindo regras de associação e classificação, agrupamentos.
- *Tipo de técnicas*: a escolha da técnica está fortemente relacionada com o tipo de conhecimento que se deseja extrair ou com os dados nos quais se aplicam tais técnicas. Entretanto, nota-se uma visão mais genérica, em que as técnicas são caracterizadas em mineração baseada em padrões, na estatística ou em teorias matemáticas.

Todavia, de maneira geral, é comum visualizar a mineração de dados de acordo com a tarefa a ser executada e então escolher o melhor conjunto de técnicas e algoritmos que promovam a solução. Entre essas técnicas, pode-se citar regras de associação, árvore de decisão, redes neurais, lógica difusa, algoritmos genéticos, estatística multivariada e mesmo consultas *ad hoc*. O presente trabalho possui foco na tarefa de agrupamentos, alinhado à fase de visualização do modelo proposto neste trabalho, e no modo como essa tarefa pode promover suporte à identificação de relações indiretas entre elementos textuais.

a.1) AGRUPAMENTOS

A tarefa de agrupamento tem como objetivo reunir itens de acordo com as suas similaridades (HAIR et al., 1998; JOHNSON; WICHERN, 1998). Em essência, visa

descobrir conjuntos de classes, isto é, encontrar conjuntos de agrupamentos (C_1, C_2, ..., C_n) que descrevam o comportamento dos dados. Esta técnica baseia-se principalmente na maximização das similaridades intra-agrupamento e na minimização das similaridades interagrupamentos (CHEN et al., 1996). Tal abordagem tem sido aplicada em diferentes áreas e tem demonstrado a sua relevância na Análise Exploratória de Dados (EAD), referenciada atualmente como Mineração de Dados (JAIN et al., 1999).

Agrupamentos possuem como base a utilização de similaridades ou distâncias entre os objetos (JOHNSON; WICHERN, 1998; GRABMEIER; RUDOLPH, 2002), visando à determinação de grupos semelhantes de objetos com o intuito de extrair conhecimentos latentes. Esse tipo de análise pode ser classificado como objetivo e procura quantificar características estruturais de um conjunto de observações (instâncias). Entretanto, por ser um processo não supervisionado, surgem desafios. Segundo HAIR et al. (1998), são três os principais desafios:

a) como avaliar a similaridade entre grupos. De modo a determinar a qualidade dos algoritmos de agrupamentos, métodos de avaliação dos resultados são extremamente relevantes. Modelos baseados no erro quadrático médio são utilizados na avaliação de dados numéricos (HALKIDI et al., 2001; DUDA; HART, 1973) e podem ser adaptados para dados textuais. A variância e a densidade de cada agrupamento são discutidas em Zaiane (2002) de modo a medir ambas as dissimilaridades interagrupamentos e similaridades intra-agrupamentos. Uma abordagem proposta por Yun et al. (2006), baseada no Ganho de Informação, avalia o resultado do processo de agrupamento sobre dados transacionais, similares aos produzidos em compras de supermercado;

b) como constituir um agrupamento. Objetiva especificar quais variáveis (dimensões) fazem parte do processo. O resultado desse processo está altamente

relacionado com o tipo de base de dados, estruturada ou não estruturada. No primeiro caso, testes estatísticos podem determinar qual dimensão fará parte do processo, o que em geral produz poucas dimensões. Por outro lado, a utilização de bases textuais (não estruturadas) pode gerar milhares de dimensões, implicando em aumento da complexidade devido principalmente à alta dimensionalidade e à esparsidade dos dados; e

c) quantos agrupamentos devem ser criados. Geralmente duas abordagens são levadas em consideração para controlar a formação dos grupos: (a) baseada em um número fixo de agrupamentos; e (b) baseada em um raio de abrangência do agrupamento. Em ambas as situações a determinação do número ideal de agrupamentos constitui-se em tarefa não trivial.

Analisando-se os itens enumerados acima, torna-se evidente a complexidade do processo. Muitas variáveis podem estar envolvidas, motivo pela qual a experiência daqueles que analisam a informação passa a ser fator indispensável. Holsheimer e Siebes (1994) demonstram que a complexidade do processo é especificada através do número de distintas formas de classificar N tuplas em k agrupamentos não vazios, e definida como $C(N) = \sum_{k=1}^{N} P(N,k) = \sum_{k=1}^{N} \frac{1}{k!} \sum_{j=0}^{k} \binom{k}{j} (-1)^j (k-j)^N$. Seja um conjunto de dados com 8 tuplas e 8 agrupamentos, 4.140 diferentes configurações (combinações) de agrupamentos serão possíveis.

a.1.1) Algoritmos

Algoritmos de agrupamentos são amplamente utilizados no estabelecimento de padrões quando existe pouca informação a respeito dos dados. Várias abordagens têm sido propostas, seja com base estatística, tais como *k*-means (MACQUEEN, 1967) e métodos Bayseanos (RAMONI et al., 2002), ou com inspiração biológica através de redes neurais, tais

como Kohonen-SOM (KOHONEN, 1995; KOHONEN et al., 2000; PANDYA et al., 1995) e ART (CARPENTER; GROSSBERG, 1987).

Entre os algoritmos não hierárquicos, o algoritmo *k*-means é um dos mais utilizados. Isso se deve basicamente pela sua facilidade de implementação (JAIN et al., 1999). Possui como principal limitação a necessidade de escolha adequada do número de agrupamentos de modo a evitar que o algoritmo atinja mínimos locais. Contudo, abordagens mais complexas têm sido propostas com o objetivo de superar limitações e implementar controles mais robustos com o intuito de evitar mínimos locais. Entre essas abordagens citam-se os trabalhos de Zha et al. (2001) utilizando decomposição QR e de Ding e He (2004) utilizando análise de componentes principais.

Nos modelos com inspiração biológica, as Redes Neurais Artificiais (ANN) têm sido amplamente utilizadas, principalmente em aplicações que exijam aprendizagem não supervisionada e/ou adaptativa. A rede neural de Kohonen-SOM (*Self-Organizing Maps*) (KOHONEN, 1995; KOHONEN et al. 2000) tem sido empregada em uma grande variedade de domínios de aplicação. Possui como característica a aprendizagem não supervisionada e elevada capacidade de generalização após a fase de treinamento. A arquitetura básica de Kohonen é composta de duas camadas completamente conectadas. A primeira representa a camada de entrada na qual a informação é apresentada à rede; e a segunda, a camada de Kohonen (**Figura 8**), realiza a classificação, podendo essa ser constituída de uma ou mais dimensões.

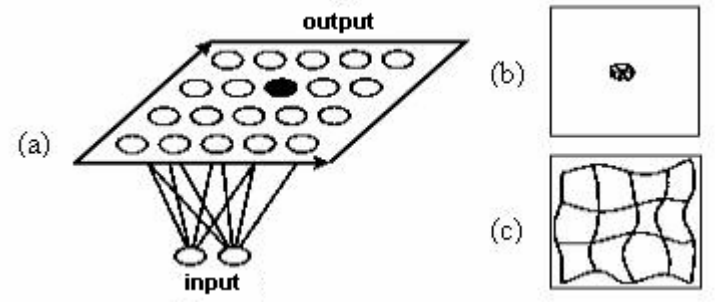


Figura 8 - (a) Rede de Kohonen com duas entradas e uma camada de aprendizado de 4x5, (b) estrutura de pesos antes da fase de treinamento e (c) estrutura de pesos após a fase de treinamento

FONTE: Adaptado de: MEDLER, 1998.

A camada de aprendizagem baseia-se na competitividade que os neurônios da camada de saída têm entre si visando estabelecer um vencedor, isto é, o neurônio mais representativo em função do padrão de entrada. Durante a fase de treinamento, o neurônio vencedor tem seus pesos atualizados, assim como os neurônios localizados na vizinhança determinada por um raio de abrangência. Nesse tipo de arquitetura, os neurônios são usualmente organizados em estruturas bidimensionais, sendo possível, entretanto, a utilização de estruturas com mais dimensões.

Ambos os modelos discutidos acima (estatístico e neural) podem ser convertidos para lidar com elementos textuais de modo a descobrir relacionamentos latentes em coleções de documentos. Contudo, devido à natureza dessa informação surgem desafios para modelos de agrupamentos, uma vez que tais algoritmos possuem em geral elevado custo computacional. Cita-se a necessidade de se lidar com centenas ou mesmo milhares de dimensões. Cada dimensão é em geral representada por termos, conceitos ou entidades que compõem o vetor que mapeia determinado documento. Como vetores possuem diferentes tamanhos, a capacidade de se trabalhar com estruturas esparsas é também requerida.

Adicionalmente, tanto a utilização de ontologias quanto a identificação de relações através de co-ocorrências objetivam incrementar a semântica latente do espaço vetorial

para melhorar o processo de agrupamento. Hotho e Stumme (2002) aprimoram o processo de agrupamento através do mapeamento de conceitos utilizando a ontologia *WordNet* (MILLER et al., 1990). Hoto et al. (2001) propõem um modelo chamado COSA (*Concept Selection and Aggregation*) que utiliza ontologia para restringir o conjunto de características dos documentos, isto é, através da agregação de conceitos para níveis mais genéricos na ontologia. Outra abordagem baseia-se no alinhamento de pares de conceitos que compartilham o mesmo termo principal (*headword*), no mesmo domínio, mas de diferentes fontes. Como resultado, são produzidos grupos de termos semanticamente relacionados (CASTILLO et al., 2003). Cita-se ainda o trabalho de Gonçalves et al. (2005), em que se utiliza de ontologia para melhorar o contexto de representação de documentos objetivando a obtenção de melhores resultados no processo de agrupamento.

Algoritmos de agrupamentos baseados em elementos textuais possuem características similares a algoritmos tradicionais de agrupamentos, por exemplo, na definição do número de agrupamentos e da estratégia de convergência utilizada. Entretanto, essa abordagem deve, por motivos de desempenho, possuir como requisito adicional a capacidade de manipular estruturas esparsas, pois vetores que representam documentos possuem tamanhos variados. Adicionalmente, medidas de similaridade capazes de incorporar as relações entre conceitos tornam-se relevantes. Gonçalves et al. (2005) consideram a medida do co-seno com a incorporação de um componente que visa melhorar a similaridade entre pares de conceitos.

Um algoritmo básico de agrupamento que, com as devidas adaptações, possibilita a utilização de estruturas textuais (vetores) em ambas as abordagens estatística e neural, pode ser implementado como se segue:

a) Passo 1: obtém o primeiro vetor de modo a criar o primeiro agrupamento.

b) Passo 2: obtém o próximo vetor e calcula a distância (similaridade) para todos os agrupamentos, visando determinar o agrupamento mais próximo.

c) Passo 3: se a similaridade para o agrupamento mais próximo for inferior a um determinado limiar (raio de abrangência), o vetor é atribuído para o agrupamento, e o centróide, recalculado. Caso contrário, se for superior a todos os agrupamentos, um novo agrupamento será criado. Se o vetor muda de um agrupamento para o outro, tanto o novo quanto o agrupamento antigo devem ter os seus pesos atualizados.

d) Passo 4: testa a convergência do algoritmo. Se essa não for atingida, retorna para o passo 2, do contrário, finaliza o algoritmo. Uma abordagem de convergência simples é a não modificação ou a modificação dentro de certa tolerância da média dos centróides na época atual em relação à época anterior. O erro quadrático também pode ser utilizado em vez da média.

2.6 CONSIDERAÇÕES FINAIS

Este capítulo discutiu as técnicas relacionadas ao trabalho advindas das áreas de extração e recuperação de informação, estatística e descoberta de conhecimento. Delinearam-se com mais ênfase os métodos de co-ocorrência, visto que esses promovem suporte ao modelo proposto (discutido no próximo capítulo) utilizado tanto na identificação de relacionamentos entre elementos textuais quanto na expansão vetorial. Além disso, realizou-se uma explanação sobre técnicas de agrupamentos. No modelo proposto, agrupamentos são utilizados para identificar relacionamentos complexos, bem como objetivam fornecer insumo para aplicações de Engenharia e Gestão do Conhecimento tais como comunidades de prática, redes sociais e localização de especialistas.

3 MODELO PROPOSTO

A inteligência é uma espécie de paladar que nos dá a capacidade de saborear idéias.
Susan Sontag

3.1 INTRODUÇÃO

O modelo proposto neste trabalho mapeia o espaço vetorial através da extração de elementos textuais (entidades, conceitos e termos) e da identificação de relacionamentos entre esses elementos (GONÇALVES et al., 2006a). Usando-se processos de extração de informação, entre eles, *Named Entity Recognition* (NER) e extração de conceitos, elementos textuais são identificados e nomeados, ou seja, atribuídos para uma determinada classe. Como exemplos de entidades citam-se áreas de pesquisa, organizações, projetos e pessoas.

Para cada documento que compõe uma coleção de documentos, os elementos textuais são extraídos e formam um vetor contendo informações necessárias para a fase de correlação. Através dessa fase torna-se possível o estabelecimento de relacionamentos diretos entre entidades, resultando em uma matriz de correlação. Nesse sentido, cada dimensão da matriz representa um vetor de contexto de determinado elemento textual. Esses relacionamentos servem de base à fase de expansão do espaço vetorial em que os elementos mais relacionados às dimensões de determinado vetor são adicionados.

Os vetores modificados são, então, usados em um processo de agrupamento, objetivando a identificação de relacionamentos indiretos estabelecidos pelas suas similaridades. Tais relacionamentos podem descrever padrões complexos de modo a serem inspecionados e utilizados em aplicações de Engenharia e Gestão do Conhecimento. Sugerem ainda, por exemplo, redes de colaboração com o objetivo de fomentar a interação e o compartilhamento de conhecimento. Por exemplo, o fato de duas entidades da classe “pessoa”

não terem relacionamento direto, ou seja, não co-ocorrerem na coleção de documentos, mas possuírem um perfil similar, possibilita em um ambiente de gestão do conhecimento a oportunidade de formação de redes de colaboração e/ou o compartilhamento de conhecimento.

A fase de validação avalia o espaço vetorial modificado pelo modelo de correlação utilizando métricas padronizadas da área de IR, como precisão, lembrança e medida *F*. Esses vetores são também avaliados através do erro quadrático em um processo de agrupamento aplicado em uma coleção de documentos. Adicionalmente, os relacionamentos diretos estabelecidos entre entidades são avaliados em três cenários por meio da precisão da correlação, incluindo comparações com julgamento de avaliadores humanos bem como uma comparação com um mecanismo de busca tradicional.

Finalmente, a fase de gestão do conhecimento utiliza-se de ferramentas de visualização gráfica dos relacionamentos estabelecidos, objetivando auxiliar, por exemplo, no entendimento do contexto organizacional.

A Figura 9 apresenta todas as fases do modelo proposto, tendo como base a correlação de elementos textuais e a expansão do espaço vetorial.

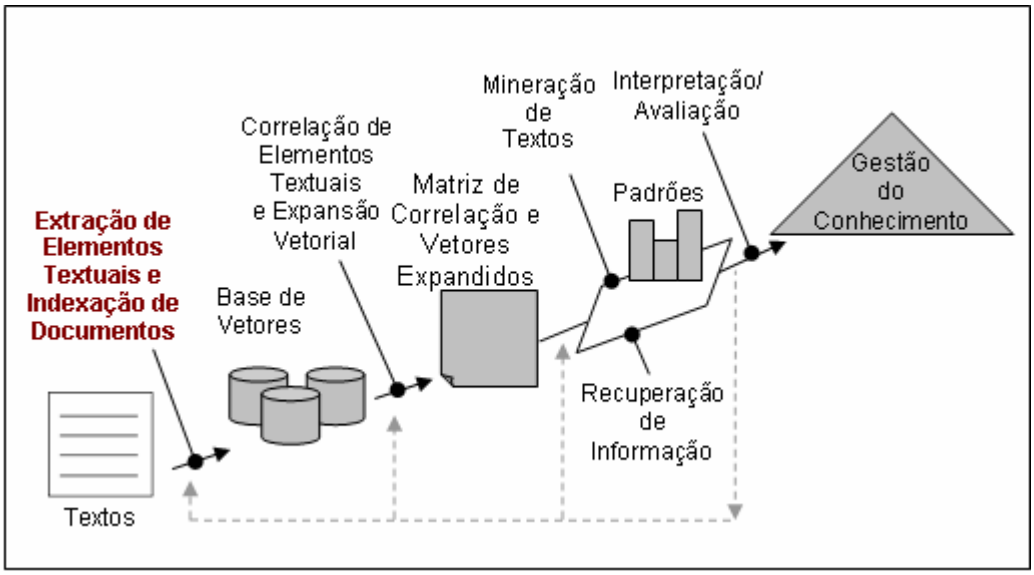


Figura 9 - Modelo de mineração de textos voltado a aplicações de Engenharia e Gestão do Conhecimento

Extração de elementos textuais: esta fase objetiva a extração de elementos textuais através de *Named Entity Recognition* (NER) e da identificação de conceitos importantes; é assim responsável por identificar, extrair e nomear elementos textuais. Entre os sistemas/arquiteturas destacam-se o sistema Espotter (ZHU et al., 2005b) e a arquitetura Gate (BONTCHEVA et al., 2004). O processo de NER constitui-se na análise de cada documento através da utilização de bases de conhecimento e de análises de padrões léxicos, possibilitando assim a extração e a nomeação de elementos textuais. Adicionalmente, análises estatísticas são aplicadas para recuperar conceitos relevantes que eventualmente não estejam mapeados na base de conhecimento. Esses conceitos são classificados para uma determinada classe geral e tratados genericamente como entidades. Informações adicionais são apresentadas na Seção 3.2.

Correlação de elementos textuais: esta é a principal fase do modelo responsável pela análise dos vetores, compostos na fase de extração, e pelo estabelecimento das relações entre os elementos textuais ao longo da coleção de documentos. Como resultado, uma matriz de correlação é estabelecida, possibilitando a determinação de relações diretas, ou seja, relações que co-ocorrem entre esses elementos nos documentos. Considerando-se, por exemplo, um par de entidades ($E1$, $E2$), a força da relação é estabelecida através da soma de todas as relações intradocumento encontradas. Informações adicionais do modelo de correlação são discutidas na Seção 3.3.

Expansão vetorial: a expansão vetorial é suportada pela matriz de correlação. Nesse sentido, para cada elemento textual que compõe o documento, os k elementos mais importantes são identificados e um vetor auxiliar é gerado. Através do vetor auxiliar, os elementos mais relevantes são adicionados ao vetor original até o limite definido no parâmetro k. A expansão vetorial objetiva um melhor mapeamento do contexto de um

determinado documento ou de qualquer unidade de análise. Informações adicionais serão apresentadas na Seção 3.4.

Geração de padrões: após a expansão dos vetores, esses são utilizados em um processo de agrupamento visando à identificação de padrões mais complexos nas relações entre os elementos textuais, ou seja, possíveis relacionamentos indiretos. Informações adicionais são exibidas na Seção 3.5.

Visualização: a fase anterior produz padrões que nem sempre são de fácil interpretação. No contexto do trabalho, a visualização desses padrões preocupa-se em como apresentar elementos textuais e suas relações, formando, assim, mapas de conhecimento. Nesse sentido, a visualização gráfica desses padrões (redes de interação entre os elementos) tem impacto na tentativa de entender como os relacionamentos, diretos e indiretos, são estabelecidos. Através da visualização espera-se obter *insights* a respeito do comportamento dessas redes para auxiliar na tomada de decisão, como, por exemplo, na definição de equipes de projetos. Informações adicionais são discutidas na Seção 3.6.

3.2 EXTRAÇÃO DE ELEMENTOS TEXTUAIS

Esta fase utiliza-se de técnicas de extração de informação na identificação de termos a partir de informação escrita que podem ou não ser nomeadas para determinada classe. O processo baseia-se, como mencionado anteriormente, na análise de texto procurando identificar elementos textuais relevantes que possam ser, em etapa posterior, classificados como uma determinada entidade. Para o presente trabalho, elementos textuais são genericamente denominados de entidades. A Figura 10 apresenta um exemplo em que os elementos textuais são identificados.

A Plataforma Lattes

A Plataforma Lattes representa a experiência do CNPq na integração de bases de dados de currículos e de instituições da área de ciência e tecnologia em um único Sistema de Informação, cuja importância atual se estende não só às atividades operacionais de fomento do CNPq como também às ações de fomento de outras agências federais e estaduais.

Dado seu grau de abrangência, as informações constantes da Plataforma Lattes podem ser utilizadas tanto no apoio a atividades de gestão como no apoio à formulação de políticas para a área de ciência e tecnologia.

O Currículo Lattes registra a vida pregressa e atual dos pesquisadores, sendo elemento indispensável à análise de mérito e competência dos pleitos apresentados à Agência.

A partir do Currículo Lattes, o CNPq desenvolveu um formato padrão para coleta de informações curriculares hoje adotado não só pela Agência mas também pela maioria das instituições de fomento, das universidades e dos institutos de pesquisa do País.

A adoção de um padrão nacional de currículos, com a riqueza de informações que esse sistema possui, a sua utilização compulsória a cada solicitação de financiamento e a disponibilização pública desses dados na internet deram maior transparência e confiabilidade às atividades de fomento da Agência.

Figura 10 - Exemplo de um documento com elementos textuais identificados durante a fase de extração de informação

Nota: texto extraído a partir da página introdutória da Plataforma Lattes[16].

Após a identificação dos elementos o processo de nomeação/classificação (NER) é suportado por uma base de conhecimento em que cada classe de entidade possui uma estrutura léxica associada. Essa estrutura é composta de um conjunto de palavras utilizadas na análise e na classificação de determinado padrão (elemento textual) como uma entidade de tipo específico. Considerando-se o documento acima (Figura 10) a base de conhecimento teria entradas (palavras) em algumas tabelas léxicas, entre elas, a tabela de projeto, a tabela de organização e uma tabela de conteúdo geral (Tabela 17).

Projeto	Organização	Geral
Plataforma, Lattes, Currículo	CNPq, Universidade, Instituto	Integração, Base, Dado, Ciência, Tecnologia, Sistema, Informação, Fomento, Agência, Federal, Estadual, Pesquisador, Competência, Currículo, Instituição, Universidade, Instituto, Padrão, Nacional, Financiamento, Transparência, Confiabilidade, Atividade

Tabela 17 - Conjunto de palavras para as tabelas léxicas de projeto, organização e geral, extraído a partir do documento apresentado na Figura 10

[16] Disponível em: <http://lattes.cnpq.br/conheca/con_apres.htm>.

Como resultado, o documento é transformado em um vetor em que cada entrada refere-se distintamente a uma entidade. Por exemplo, a entidade “CNPq” atribuída para a classe “Organização” foi mencionada três vezes em diferentes partes do texto. A Tabela 18 apresenta a estrutura do vetor de entidades.

```
<entities>
        <entity name=”CNPq” class=”Organization” pos=”11;48;126>”/>
        <entity name=”Plataforma Lattes” class=”Projeto” pos=”2;5;70”/>
        <entity name=”Currículo Lattes” class=”Projeto” pos=”97;123”/>
        <entity name=”Ciência e Tecnologia” class=”Geral” pos=”26;93”/>
        ….
</entities>
```

Tabela 18 - Exemplo do vetor de entidades gerado a partir do documento apresentado na Figura 10

3.3 CORRELAÇÃO DE ELEMENTOS TEXTUAIS

A correlação de elementos textuais tem sido utilizada na identificação de seqüências naturais de palavras que possivelmente indicam conceitos a serem extraídos a partir de informação escrita. Adicionalmente, tais métodos podem ser aplicados a qualquer tipo de elemento textual para indicar proximidade ou força de relacionamento que esses elementos mantêm entre si.

Para o processo de correlação o conjunto de vetores de entidade representando a coleção de documentos (sendo um vetor em particular similar ao apresentado na Tabela 18) pode ser representado como uma matriz entidade–documento. Como exemplo são considerados três documentos que possuem ao todo sete entidades. Cada entidade em um determinado documento ocorre em *n* posições, como demonstrado na Tabela 19.

Entidades	***D*1**	***D*2**	***D*3**
*E*1	<10, 100, 180, 350>	<50, 110>	<>
*E*2	<20, 140>	<>	<20, 100, 170>
*E*3	<195, 270, 310>	<80, 150>	<>
*E*4	<210>	<130>	<>
*E*5	<>	<90, 200>	<>
*E*6	<>	<>	<40, 80>
*E*7	<80>	<>	<120, 210>

Tabela 19 - Exemplo de uma matriz entidade–documento com as respectivas posições

Considerando-se a entidade (*E*1) que ocorre em diferentes documentos, existe um número de entidades que co-ocorrem com essa entidade. Com o objetivo de aprimorar o estabelecimento da força dos relacionamentos entre entidades, foi proposto o modelo de correlação LRD (*Latent Relation Discovery*). Nesse sentido, dois aspectos - co-ocorrência e distância - são considerados na determinação do grau de relacionamento entre duas entidades.

Co-ocorrência: duas entidades são consideradas co-ocorrências se elas aparecem no mesmo documento. Se uma entidade é altamente relacionada à outra entidade, geralmente elas tendem a co-ocorrerem mais freqüentemente. De modo a normalizar as ocorrências entre duas entidades, *E*1 e *E*2, a freqüência relativa (RESNIK, 1999) é utilizada, sendo definida como:

$$\hat{p}(E1, E2) = \frac{Num(E1, E2)}{N}, \tag{16}$$

onde *Num*(*E*1,*E*2) indica o número de documentos em que *E*1 e *E*2 co-ocorrem e *N* informa o número total de documentos.

Distância: duas entidades altamente relacionadas tendem a ocorrer com certa freqüência. Se duas entidades, *E*1 e *E*2, ocorrem somente uma vez no documento, a distância entre *E*1 e *E*2 é a raiz do módulo da diferença entre as posições. Se *E*1 ocorre somente uma vez e *E*2 múltiplas vezes no documento, a distância de *E*1 para *E*2 é a diferença da posição de *E*1 e a posição mais próxima do vetor de ocorrências de *E*2. Quando ambas *E*1 e *E*2 ocorrem múltiplas vezes no documento, a distância será o somatório de cada ocorrência de *E*1 para a mínima distância em relação às ocorrências de *E*2 no *i*th documento. Em ambos os casos a raiz quadrada é aplicada ao somatório das distâncias mínimas, normalizada pela freqüência de

$E1$. A equação de distância é definida como:

$$d_i(E1,E2) = \frac{\sqrt{\sum_{j=1}^{f_i(E1)} \min\left(\left|E1_j, E2\right|\right)}}{f_i(E1)}, \quad (17)$$

onde $f_i(E1)$ é o número de ocorrências de $E1$ no ith documento e $\min(|E1_j, E2|)$ é a mínima distância entre a jth ocorrência de $E1$, $E1_j$ e todas as ocorrências de $E2$. Quando a quantidade de ocorrência difere entre os vetores de posições, $E1$ e $E2$, realiza-se um alinhamento entre as posições de menor distância. Outra solução possível quando existem múltiplas ocorrências nos vetores, $E1$ e $E2$, seria o cálculo da distância em ambas as direções e o estabelecimento da média entre $d_i(E1, E2)$ e $d_i(E2, E1)$. Contudo, eventualmente uma ou mais posições do vetor $E1$ podem estar relativamente distantes da sua posição mais próxima em $E2$. Isso promove, em muitos casos, um aumento na distância intradocumento entre as entidades comparadas, o que interfere de maneira negativa no peso da relação.

Sejam $E1$={10,100,180,350} e $E2$={20,140} os vetores das entidades $E1$ e $E2$ com suas respectivas ocorrências no ith documento (neste caso o documento $D1$), a distância em ambas as direções sem o alinhamento seria $d_i(E1,E2) = \sqrt{\sqrt{(10-20)^2} + \sqrt{(100-140)^2} + \sqrt{(180-140)^2} + \sqrt{(350-140)^2}}/4 = 4{,}3301$ e $d_i(E2,E1) = \sqrt{\sqrt{(20-10)^2} + \sqrt{(140-100)^2}}/2 = 3{,}5355$, sendo a média 3,9328. Considere outro exemplo, sendo $E1$={10,100,250} e $E2$={20,140}, a distância seria $d_i(E1,E2) = \sqrt{\sqrt{(10-20)^2} + \sqrt{(100-140)^2} + \sqrt{(250-140)^2}}/3 = 4{,}2164$ e $d_i(E2,E1) = \sqrt{\sqrt{(20-10)^2} + \sqrt{(140-100)^2}}/2 = 3{,}5355$, e a média seria 3,8760. Entretanto, utilizando-se o alinhamento, ou seja, somente as posições de mínima distância dos pares de entidades em cada vetor, o resultado seria $\sqrt{\sqrt{(10-20)^2} + \sqrt{(100-140)^2}}/2 = 3.535$. A utilização somente dos pares alinhados tende a produzir resultados mais consistentes na distância intradocumento.

Grau de relacionamento: seja uma entidade $E1$, o grau ou a força de relacionamento entre duas entidades $E1$ e $E2$ é calculado levando-se em conta o número de co-ocorrências no ith documento, a distância entre as múltiplas ocorrências no ith documento e a freqüência normalizada, sendo essa equação definida como:

$$R(E1,E2) = \hat{p}(E1,E2) \times \sum_{i} \left(\frac{f(Freq_i(E1)) \times f(Freq_i(E2))}{d_i(E1,E2)} \right), \tag{18}$$

onde $f(Freq_i(E1)) = tfidf_i(E1)$, $f(Freq_i(E2)) = tfidf_i(E2)$, e $Freq_i(E1)$ e $Freq_i(E2)$ são o número de ocorrências de $E1$ e $E2$ no ith documento, respectivamente. A medida de normalização *tfidf* é definida como $tfidf(ij) = tfe_{ij} * \log_2 \left(N / dfe_i \right)$, onde $tfe_{ij} = fe_{ij} / max(fe_{ij})$ é a freqüência (*fe*) normalizada da entidade e_j no documento j, N é o número total de documentos no coleção e dfe_i é o número de documentos que contêm a entidade e_i.

Analisando-se a equação, observa-se que quanto maior a distância entre duas entidades menor será o grau de relacionamento entre elas, e vice-versa.

3.4 EXPANSÃO DO ESPAÇO VETORIAL

A fase de composição/expansão vetorial tem como objetivo incrementar o contexto do espaço vetorial de modo a criar uma representação mais complexa que possibilite suporte para atividades de recuperação de informação, descoberta de conhecimento e gestão do conhecimento. De modo geral, assume-se que existe uma estrutura semântica latente que quando identificada pode incrementar a representação de documentos. Isso ocorre através da localização de termos relevantes aos termos originalmente constantes no documento. Por exemplo, na recuperação de informação a redefinição do contexto permite a localização de documentos, que, caso fossem utilizados os documentos originais, não seriam apresentados ao usuário.

Para um melhor entendimento do processo, considere a matriz entidade–documento (Tabela 20) composta de três colunas originais (*D*1, *D*2, e *D*3) com as freqüências absolutas de cada entidade em cada documento (sumarizada a partir da Tabela 19), e de três colunas normalizadas (*D*1-N, *D*2-N e *D*3-N) pela medida *tfidf*.

Entidades	***D*1**	***D*2**	***D*3**	***D*1-N**	***D*2-N**	***D*3-N**
*E*1	4	2	0	0,5850	0,5850	0
*E*2	2	0	3	0,2925	0	0,5850
*E*3	3	2	0	0,4387	0,5850	0
*E*4	1	1	0	0,1462	0,2925	0
*E*5	0	2	0	0	1,5850	0
*E*6	0	0	2	0	0	1,0566
*E*7	1	0	2	0,1462	0	0,3900

Tabela 20 - Exemplo de uma matriz entidade–documento

Nota: a segunda parte da matriz, identificada pelas colunas *D*1-N, *D*2-N e *D*3-N, indica os valores normalizados pela equação *tfidf*.

Seja um documento *D*1, o conjunto de relações entre entidades é estabelecido analisando-se todas as possibilidades distintas entre entidades que possuam ocorrência no documento, não importando a ordem no par. Por exemplo, tomando-se como entrada a entidade *E*3, as relações seriam (*E*1,*E*3), (*E*2,*E*3), (*E*3,*E*4) e (*E*3, *E*7). Para cada relação, são disponibilizadas as freqüências das entidades de origem e destino, a distância da relação e o peso da relação (peso parcial). Esses valores são armazenados como um índice invertido bidirecional, similar ao empregado em sistemas de recuperação de informação, em que cada entrada na tabela léxica (representada pelos quadrados na Figura 11) possui uma determinada entidade e um ponteiro para as demais informações relacionadas.

Seja *E*1 uma entrada na tabela léxica, existe um apontamento para as relações com *E*2, *E*3, *E*4, *E*5, *E*7, e seus respectivos valores. Como uma entidade pode aparecer tanto na origem quanto no destino, os apontamentos na estrutura de armazenamento são bidirecionais. Do mesmo modo a entidade *E*3 também possui ponteiros para todas as entidades com as quais ela se conecta. Tal estrutura facilita a rápida identificação de listas ordenadas de entidades relacionadas a partir de uma entidade de origem.

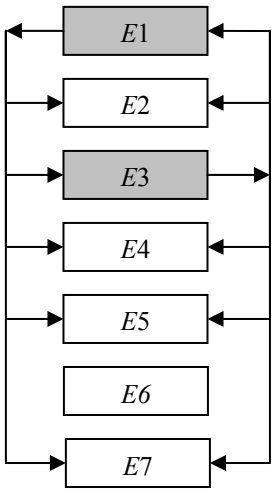


Figura 11 - Representação da estrutura de armazenamento das entidades e suas relações

A Tabela 21 apresenta os pares de entidades relacionadas (ERs) para os três documentos anteriores (Tabela 20), as freqüências, as distâncias intradocumento calculadas através da Equação 17 e o grau de relacionamento intradocumento calculado através da segunda parte da Equação 18, ou seja, $(f(Freq_i(E1)) \times f(Freq_i(E2))) / d_i(E1, E2))$.

Id do Documento	Origem	Freq.	Destino	Freq.	Distância	Peso Parcial
1	E1	4	E2	2	3,5355	0,0484
1	E1	4	E3	3	3,8730	0,0663
1	E1	4	E4	1	5,4772	0,0156
1	E1	4	E7	1	4,4721	0,0191
1	E2	2	E3	3	7,5829	0,0169
1	E2	2	E4	1	8,3666	0,0051
1	E2	2	E7	1	7,7460	0,0055
1	E3	3	E4	1	3,8730	0,0166
1	E3	3	E7	1	10,7238	0,0060
1	E4	1	E7	1	11,4018	0,0019
2	E1	2	E3	2	3,8730	0,0884
2	E1	2	E4	1	4,4721	0,0383
2	E1	2	E5	2	3,8730	0,2394
2	E3	2	E4	1	4,4721	0,0383
2	E3	2	E5	2	3,8730	0,2394
2	E4	1	E5	2	6,3246	0,0733
3	E2	3	E6	2	3,1623	0,1955
3	E2	3	E7	2	3,8730	0,0589
3	E6	2	E7	2	5,4772	0,0752

Tabela 21 - Tabela de documentos e co-ocorrência de entidades com os pesos intradocumento para cada relação

Através da Tabela 21 é possível calcular as correlações entre as entidades extraídas da coleção de documentos. Seja um par de entidades (*E*1, *E*3), o grau de relacionamento é calculado através da Equação 18, $R(E1, E3) = 2/3 \times (0{,}0663 + 0{,}0884) = 0{,}1031$, e apresentado na Tabela 22.

Origem/Destino	***E*1**	***E*2**	***E*3**	***E*4**	***E*5**	***E*6**	***E*7**
***E*1**	-	0,0161	0,1031	0,0359	0,0798	0	0,0064
***E*2**	0,0161	-	0,0056	0,0017	0	0,0652	0,0429
***E*3**	0,1031	0,0056	-	0,0365	0,0798	0	0,0020
***E*4**	0,0359	0,0017	0,0365	-	0,0244	0	0,0006
***E*5**	0,0798	0	0,0798	0,0244	-	0	0
***E*6**	0	0,0652	0	0	0	-	0,0251
***E*7**	0,0064	0,0429	0,0020	0,0006	0	0,0251	-

Tabela 22 - Tabela de adjacência (matriz de correlação) apresentando o grau de relacionamento entre as entidades da coleção de documentos

A tabela de adjacência (matriz de correlação) é utilizada no processo de expansão vetorial. Para cada dimensão (entidade) são identificadas as *k* mais importantes entidades, ou seja, que possuem os maiores níveis de relacionamento, e são adicionadas aquelas que não fazem parte do vetor original que representa o documento. Para a expansão do documento *D*1 verifica-se que as entidades *E*5 e *E*6 não fazem parte do vetor. Analisando-se a Tabela 22 e utilizando-se *k*=2, as entidades *E*5 com valor de 0.0798 e *E*6 com valor de 0.0652 serão selecionadas através da relação com as entidades *E*3 e *E*2, respectivamente. Considerando-se *D*2, as entidades *E*2 e *E*7 serão adicionadas através da relação com *E*1 em ambos os casos. Finalmente, as entidades *E*1 e *E*3 serão adicionadas ao documento *D*3 através da relação com *E*2.

Para o cálculo final dos pesos das entidades adicionadas ao vetor original em cada documento, é novamente utilizada a Tabela 22. O peso do novo elemento adicionado $w(E_{novo})$, é definido como:

$$w(E_{novo}) = \sum_{i=1}^{num(E_{novo}, D)} R(E_{novo}, E_i) \times w(E_i), \qquad (19)$$

onde $R(E_{novo},E_i)$ é a força do relacionamento entre E_{novo} e E_i, que está na matriz de correlação, $w(E_i)$ é o peso de E_i no documento D e $num(E_{novo},D)$ é o número total de entidades originalmente no vetor do documento que se relacionam com E_{novo}. Por exemplo, o peso de *E*5 em *D*1 será: 0,5850*0,0798+0,4387*0,0798+0,1462*0,0244 = 0,0853. A Tabela 23 apresenta os vetores expandidos para os três documentos, *D*1, *D*2 e *D*3, utilizando *k*=2.

Entidades	***D*1**	***D*2**	***D*3**
*E*1	0,5850	0,5850	0,0119
*E*2	0,2925	0,0132	0,5850
*E*3	0,4387	0,5850	0,0041
*E*4	0,1462	0,2925	0
*E*5	0,0853	1,5850	0
*E*6	0,0227	0	1,0566
*E*7	0,1462	0,0051	0,3900

Tabela 23 - Exemplo de uma matriz entidade–documento com os vetores expandidos

3.5 GERAÇÃO DE PADRÕES

A fase de geração de padrões tem como objetivo permitir análises mais elaboradas sobre a coleção de documento, visando à identificação de relacionamentos indiretos entre entidades.

Para tal, um algoritmo de agrupamento baseado no algoritmo *k*-means (MACQUEEN, 1967) tem sido utilizado[17] (GONÇALVES et al., 2005). Contudo, ao contrário da versão tradicional, que considera um número fixo de agrupamentos, a versão aqui proposta utiliza um raio de abrangência (*r*). Nesse sentido, os objetos que estiverem dentro desse raio serão classificados como similares, do contrário, dissimilares. O algoritmo possui os seguintes passos:

- passo 1: selecionar o primeiro vetor da lista de vetores. Caso ainda não existam agrupamentos, deve-se criar um agrupamento com todas as dimensões (entidades);
- passo 2: selecionar o próximo vetor;

[17] Considerando-se que o foco do trabalho consiste na avaliação da correlação de elementos textuais e na expansão vetorial de documentos o algoritmo *k*-means foi utilizado devido à sua simplicidade de implementação.

- passo 3: determinar através da equação do co-seno (Equação 1) qual o agrupamento mais similar. Se a distância d=1-cosθ for menor que o raio (r), os pesos do agrupamento serão atualizados recalculando-se os centros; do contrário, um novo agrupamento será criado. Se o vetor que está sendo analisado for alterado de um agrupamento para outro, ambos os agrupamentos, antigo e novo, devem ter seus centros recalculados; e
- passo 4: repetir os passos 2 e 3 até que não haja diferenças na média total dos centróides ou erro quadrático entre a iteração atual e a anterior ou que um determinado número de épocas[18] seja atingido.

A utilização de um algoritmo de agrupamento é parte importante no modelo, pois, como mencionado, objetiva a identificação de padrões não triviais e indiretos. Contudo, a utilização dessa classe de algoritmo exige a definição de parâmetros de execução, uma tarefa nem sempre fácil que em geral dependente da natureza dos dados. Discussões sobre os parâmetros utilizados na execução do algoritmo são apresentadas na Seção 4.

3.6 VISUALIZAÇÃO DE PADRÕES

A interpretação de resultados produzidos no processo de agrupamentos nem sempre é uma tarefa fácil ou intuitiva. Diferentes modelos e ferramentas têm sido propostos. Análises de dados podem ser facilitadas através de sumarizações por meio de média, desvio padrão e variância, por exemplo. Entretanto, isso depende do tipo de dado a ser analisado. No caso de elementos textuais (entidades), além desses recursos, a visualização gráfica possibilita a interconexão entre eles, provendo meios para facilitar o entendimento do conhecimento latente em bases textuais. Os elementos textuais conectados formam assim um mapa do conhecimento e promovem suporte às aplicações de gestão do conhecimento, tais como

[18] Por "época" entende-se como uma varredura executada sobre todos os vetores selecionados para o processo de agrupamento.

comunidades de prática, redes sociais e localização de especialistas. O entendimento dessas estruturas interconectadas pode auxiliar na definição de estratégias e na tomada de decisão de modo a otimizar recursos nas organizações.

A visualização do modelo utiliza-se do resultado produzido na fase de geração de padrões e é suportada pela tabela de relacionamentos diretos (Tabela 22). Essa tabela permite, através da indicação de uma entidade ou de um par de entidades, a obtenção da lista de entidades relacionadas com os seus respectivos pesos ou simplesmente com o peso da relação quando informado somente um par de elementos. Apesar de cada célula considerar apenas uma simples relação, analisando-se toda a estrutura, torna-se possível o estabelecimento de caminhos complexos que conectam entidades sem necessariamente essas entidades estarem diretamente relacionadas. Sendo assim, através da base de relações (matriz de correlação), as conexões entre entidades que foram reunidas no mesmo agrupamento são estabelecidas.

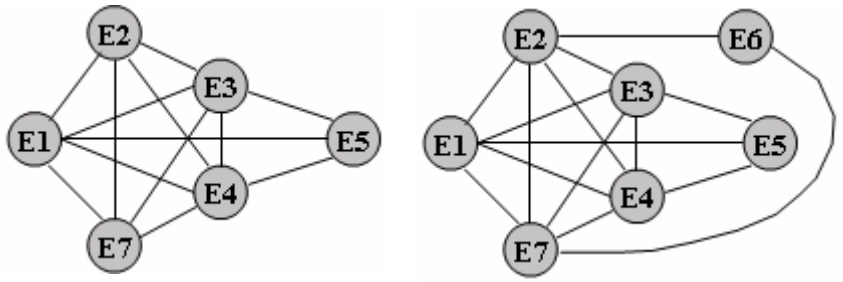
Ainda assim identificar e explicar o motivo pelo qual tais entidades estão interconectadas pode ser uma tarefa de difícil execução, uma vez que para isso a extração dos tipos das relações entre entidades seria de grande utilidade, embora de difícil implementação. Este trabalho está limitado a somente apresentar as entidades interconectadas e a força dessas relações de modo a auxiliar em tarefas de gestão do conhecimento.

Para exemplificar a proposta, a Tabela 20 (colunas *D*1-N, *D*2-N e *D*3-N) e a Tabela 23 são consideradas como entradas para dois processos de agrupamentos, e a Tabela 22 é considerada como insumo para realizar a ligação entre as entidades. Como resultado do processo, dois grupos serão criados para cada uma das tabelas. Em ambos os casos, o primeiro agrupamento reúne os documentos *D*1 e *D*2, enquanto que o segundo agrupamento possui o documento *D*3. A Figura 12a apresenta o primeiro agrupamento utilizando como fonte a Tabela 20. Todas as entidades são unidas através da tabela de relacionamentos (Tabela 22). A Figura 12b também apresenta o primeiro agrupamento, mas considerando a Tabela 23, com os

vetores expandidos. Novamente as entidades são unidas por meio da tabela de relacionamentos.

Como resultado do processo de agrupamento, utilizando-se dos vetores expandidos, a entidade *E*6 é adicionada ao primeiro agrupamento. Pesquisando-se a entidade *E*6 na tabela de relacionamentos, verifica-se que as entidades *E*2 e *E*7 podem ser interconectadas.

Desse modo, a tabela de relacionamentos diretos promove os meios para o estabelecimento de relações entre entidades, permitindo assim a visualização, a inspeção e a análise de cenários mais complexos (Figura 12).



(a) (b)

Figura 12 - Gráfico de entidades interconectadas

Nota: (a) utilizando o espaço vetorial tradicional e (b) utilizando a abordagem de espaço vetorial expandido. Cada figura apresenta um agrupamento diferindo no número de entidades. Como pode ser observado, na Figura b a entidade *E*6 é adicionada como resultado do novo espaço vetorial. Para efetuar as conexões, a tabela de relacionamentos diretos é utilizada.

3.7 CONSIDERAÇÕES FINAIS

Este capítulo apresentou o modelo proposto, detalhando cada uma das fases que o constitui. O núcleo do modelo é composto de um método de correlação de elementos textuais e de um modelo de expansão vetorial, chamado *Latent Relation Discovery* (LRD). Adicionalmente, o modelo incluiu outras fases para que se atinjam os objetivos do trabalho, entre elas, a extração de elementos textuais, a indexação de unidades de análises, a aplicação de técnicas de mineração de textos e a visualização dos mapas de conhecimento. Nesse

sentido, a integração dessas técnicas em um único modelo visa fornecer um ferramental que promova suporte às aplicações de Engenharia e Gestão do Conhecimento.

4 APRESENTAÇÃO DOS RESULTADOS

Há três séculos, o conhecimento científico não faz mais do que provar suas virtudes de verificação e de descoberta em relação a todos os outros modos de conhecimento.

Edgar Morin

4.1 INTRODUÇÃO

O modelo proposto neste trabalho apresenta problemas típicos de validação de mecanismos de aprendizagem não supervisionados devido, principalmente, à falta de parâmetros de comparação para verificar se o modelo foi capaz de aprender ou cumprir com o seu propósito. As principais abordagens para a avaliação podem ser caracterizadas como quantitativas, conjuntos de dados de avaliação (*gold standard*) ou orientadas a tarefas, cuja descrição é apresentada na seqüência.

Quantitativa: métodos desta natureza julgam se o resultado ou o modelo produzido é adequado segundo parâmetros de confiabilidade. Por exemplo, um método clássico para análise de algoritmos de agrupamentos hierárquicos aglomerativos é o coeficiente de correlação *cophenetic* (SOKAL; ROHLF, 1962; HALKIDI et al., 2001). O erro quadrático é freqüentemente utilizado para avaliar a eficiência de modelos de agrupamentos sobre dados numéricos (DUDA; HART, 1973). Outro modelo recente utiliza-se do Ganho de Informação na avaliação da qualidade da tarefa de agrupamento sobre dados transacionais, similar aos gerados em compras de supermercados (YUN et al., 2006).

Conjuntos de dados de avaliação (*Gold standard*): esta abordagem compara o modelo aprendido com um modelo "ideal" produzido *a priori* por um especialista no domínio. Isso é comum na área de recuperação de informação e classificação de documentos, como,

por exemplo, a série de competições MUC (DARPA, 1995), TREC[19], as bases que compõem o projeto SMART[20] e a base REUTERS[20]. A principal desvantagem reside no fato de coleções padronizadas possuírem custo elevado para a sua produção. Apesar de essas bases de dados serem produzidas por especialistas, a subjetividade é também um problema. Por outro lado, uma vez produzidas, essas bases podem ser aperfeiçoadas com o tempo e passam a servir de referência na avaliação e na comparação de diversos modelos e algoritmos.

Orientada a tarefas: esta abordagem examina a utilidade de modelos considerando ambientes realistas, com monitoramento científico e objetivo das atividades em andamento. É obtida através da utilização de diferentes grupos de usuários, que executam tarefas do mundo real e sob condições e ambiente monitorados. A subjetividade, assim como no modelo anterior, faz parte do processo. Tonella et al. (2003) discutem alguns dos problemas associados a esta abordagem, incluindo o seu custo e a necessidade de um projeto cuidadoso para minimizar a subjetividade. Os autores citam ainda que o custo elevado deve-se principalmente à necessidade de trabalho humano intensivo na definição e na execução das tarefas de avaliação e validação.

O cerne do modelo proposto é composto da correlação de elementos textuais e expansão vetorial. Considerando-se a discussão acima, a abordagem orientada a tarefas seria adequada desde que fosse possível produzir um ambiente controlado para avaliação e validação do modelo. Pelos motivos também já citados, o custo e a subjetividade podem introduzir erros no processo de avaliação. Desse modo, a validação do trabalho baseia-se nas três abordagens discutidas anteriormente, mas com foco na avaliação utilizando conjuntos de dados padronizados (*gold standards*) e quantitativos, e buscando assim definir um modelo automático de validação.

[19] Disponível em: <http://trec.nist.gov/>.
[20] Disponível em: <http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/>.

Para a abordagem *gold standard* dois conjuntos de documentos foram considerados. O primeiro conjunto é utilizado na avaliação do modelo quando aplicado a sistemas de recuperação de informação, enquanto que o segundo, na avaliação do modelo quando aplicado à tarefa de agrupamento de textos. Nesses conjuntos de dados, os termos (palavras) que compõem o vetor de cada documento são tratados como elementos textuais.

Em ambos os casos, os vetores que representam cada documento dessas bases foram expandidos através do modelo de correlação, ou seja, os *k* termos mais relevantes em relação aos termos originais do vetor são utilizados para expandir o espaço vetorial. Os termos mais relevantes são definidos em função do método de correlação de elementos textuais, sendo este comparado com outros métodos similares.

O processo de expansão vetorial utilizado nesses dois conjuntos de dados que compõem a validação segue os seguintes passos descritos abaixo:

1. Para cada documento é gerado o vetor de termos. Cada termo, representando uma dimensão, indica um possível elemento textual.
2. Utilizando todos os vetores da base, o modelo de correlação é executado de modo a estabelecer as relações entre os termos. Como resultado, um termo $T1$ pode relacionar-se a múltiplos termos $T1=<T2,T3,..,Tn>$, sendo armazenado conforme a Tabela 21.
3. A tabela de termos correlacionados é então utilizada para expandir os vetores que representam a coleção de documentos. Cada vetor é expandido até um determinado limiar *k,* conforme exemplo apresentado na Seção 3.4. Os valores de *k* utilizados na validação foram 5, 10, 15, 20, 25, 30, 35, 40, 45 e 50.
4. Cada vetor sofre a expansão considerando, além do método LRD, outros métodos, tais como, *Phi-squared*, MI, VMI, *Z score* e LSI. Para cada método foram consideradas as janelas de 0 (sem janela, ou seja, todo o documento é

considerado), 20, 50, 100 e 200. A janela indica o deslocamento à direita ou à esquerda de determinado termo. Por exemplo, se a posição de um termo *T*2 está dentro de uma determinada janela *x* em relação à posição do termo *T*1, a relação entre ambos os termos será considerada no cálculo da correlação. Para os métodos comparados, o conceito de janela somente não se aplicará ao LSI por não suportar essa característica. LSI utiliza as freqüências absolutas de cada termo nos seus documentos para determinar as correlações, e não as freqüências conjuntas de determinado par de termos, por meio do qual seria possível determinar a distância através de valores predeterminados de janelas. Os testes com LSI restringem-se à abordagem sem janela.

Além das duas etapas da proposta de validação automática baseada em *gold standards*, uma com foco na recuperação de informação e outra no agrupamento de textos, propõe-se mais dois estudos. O primeiro orientado à tarefa, constituído de dois estudos de caso em que usuários foram convidados a avaliar os relacionamentos entre entidades, e um outro, quantitativo, com vistas a confrontar a precisão de correlação entre elementos textuais (entidades) utilizando um mecanismo de busca tradicional como ponto de referência. Todas as abordagens são discutidas a seguir.

4.1.1 RECUPERAÇÃO DE INFORMAÇÃO

A avaliação do modelo foi realizada através das medidas de precisão, lembrança e medida *F*, aplicadas em uma base de dados padronizada utilizada na avaliação do desempenho de sistemas de recuperação de informação. Essa base, chamada de CISI[21], possui um total de 1.460 documentos e 112 consultas. Para cada consulta existe um conjunto de respostas corretas, ou seja, os documentos que a satisfazem.

[21] Disponível em: <http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cisi/>.

a) MODELO DE VALIDAÇÃO

Para se validarem cenários de recuperação de informação torna-se necessária a análise do desempenho do sistema quanto à sua capacidade em recuperar documentos relevantes. Por "documentos relevantes" entende-se aqueles que podem satisfazer determinada consulta efetuada pelo usuário. A matriz de contingência abaixo (Tabela 24) demonstra a distribuição conjunta de duas variáveis:

	relevante	¬ relevante
selecionado	*pv*	*fp*
¬ selecionado	*fn*	*pn*

Tabela 24 - Matriz de contingência utilizada no cálculo das medidas de precisão e lembrança

O número em cada célula representa as freqüências ou as contagens de documentos em cada região. Os casos encontrados para *pv* (positivos verdadeiros) e *pn* (positivos negativos) são os casos em que o sistema obtém sucesso. Os casos selecionados erroneamente em *fp* e *fn* são chamados de falsos positivos e falsos negativos, respectivamente.

A precisão é definida como uma medida de proporção de itens que o sistema selecionou corretamente. Por exemplo, em uma determinada consulta que retorne 100 documentos dos quais 60 são relevantes e 40 irrelevantes, a precisão seria de 0.6. A equação é definida como:

$$precisão = \frac{pv}{pv + fp}, \tag{20}$$

A lembrança é definida como a proporção de itens corretos que o sistema selecionou. Por exemplo, para uma determinada consulta que retorne 60 documentos corretos, deixando de recuperar outros 15 corretos, o valor seria de 0.8. A equação é definida como:

$$recall = \frac{pv}{pv + fn}, \tag{21}$$

Essas duas medidas possuem comportamento oposto, em que, por exemplo, o aumento da precisão tende a diminuir a lembrança, e vice-versa. Elas podem ainda ser combinadas objetivando obter uma simples medida do desempenho geral do sistema. Uma maneira de se

calcular isso é através da medida *F* desenvolvida por Van Rijsbergen (1979), sendo definida como:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}, \tag{22}$$

onde *P* é a precisão, *R* é a lembrança e α é o fator que determina o peso de ambas, precisão e lembrança. O valor 0.5 é freqüentemente utilizado e indica relevância proporcional para as duas medidas. Sendo assim, a medida *F* pode ser simplificada como:

$$F = \frac{2PR}{(R+P)} \tag{23}$$

b) DISCUSSÃO DOS RESULTADOS

Considerando-se os 10 valores de *k* utilizados na expansão vetorial em cinco diferentes configurações de janela, o total de conjuntos de dados para a fase de recuperação de informação é 50.

A expansão vetorial, como mencionado anteriormente, foi realizada utilizando-se a base CISI, composta de 1.460 documentos e 112 consultas. Dessas, foram selecionadas, de maneira aleatória, 20 consultas. Dada uma consulta, a medida do co-seno é calculada entre o vetor de consulta e os *n* vetores recuperados que representam os documentos. Para o limiar do co-seno, foram utilizados os valores 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 e 0.8. Sendo assim, para uma determinada consulta, a similaridade entre o vetor de consulta e os documentos é realizada empregando-se os vários limiares, uma determinada janela e um determinado valor *k*. Utilizando-se os limiares, é possível obter para a corrente consulta as medidas de desempenho (precisão, lembrança e medida *F*).

A medida *F* média para o modelo de espaço vetorial original, isto é, sem a expansão vetorial, sem janela e considerando todos os limiares de 0.1 até 0.8, é de 0,092. Essa medida promove a base para a comparação com os demais métodos aplicados na expansão dos

vetores. A Tabela 25 apresenta os valores médios (entre 0 e 1) da medida *F,* considerando todos os limiares (entre 0.1 e 0.8) e os cinco métodos de correlação, mais LSI para a configuração sem janela. LSI é o único método avaliado que é baseado no uso de documento completo em vez de uma janela. Considerando-se LSI e os outros cinco métodos aplicados em diferentes configurações de janelas, LRD teve o melhor desempenho. O melhor valor da medida *F* média é 0,1438 para janela de 20 e *k*=40. O segundo método que melhor teve desempenho foi LSI com 0,1310 para *k*=30, 9,7% abaixo da melhor medida *F* de LRD. O terceiro melhor é o *Z score,* sendo a melhor medida *F* 0,1295 para as janelas de 50 e 100 e *k*=50. Em seguida, está o método *Phi-squared,* sendo o melhor resultado 0,1096 para a janela de 200 e *k*=5. Os métodos VMI e MI possuem os piores desempenhos, mas similares entre si.

A Figura 13 demonstra a influência do fator *k* sobre os métodos utilizados. Para o método LSI, nota-se que a medida *F* se mantém praticamente estável. Para uma determinada configuração da janela, quando o fator *k* aumenta, a medida *F* dos métodos LRD e *Z score* aumenta. Contudo, no caso de LRD ocorre uma queda de desempenho a partir de *k*=45. Para a medida *F* do método LRD, o maior incremento acontece variando-se o *k* de 10 para 15. Entretanto, conseqüentes incrementos em *k* não demonstram ganhos significativos para o método LRD. O método *Z score* possui uma tendência de queda de *k*=10 para *k*=20 e depois um incremento à medida que *k* varia. Para os demais métodos, conforme *k* varia, a medida *F* decresce, com exceção do método VMI. Este possui um incremento em *k*=25. Tal fato não parece indicar que esse seja o melhor ponto de parada da expansão, uma vez que em fatores subseqüentes ocorrem novamente quedas.

Medida *F* (média)		*k*=5	*k=10*	*k*=15	*k=20*	*k*=25	*k=30*	*k*=35	*k=40*	*k*=45	*k=50*	*Média*
Sem Janela	LRD	0,1357	0,1332	0,1364	0,1371	0,1392	0,1399	0,1406	0,1411	**0,1427**	0,1406	0,1387
	LSI	0,1243	0,1260	0,1208	0,1269	0,1295	**0,1310**	0,1294	0,1248	0,1226	0,1221	0,1257
	Z Score	**0,1065**	0,1025	0,0963	0,0968	0,0978	0,0969	0,0964	0,0948	0,0954	0,0951	0,0979
	Phi-squared	**0,0989**	0,0975	0,0898	0,0911	0,0908	0,0884	0,0898	0,0922	0,0934	0,0923	0,0924
	VMI	**0,0753**	0,0675	0,0646	0,0620	0,0628	0,0617	0,0608	0,0595	0,0591	0,0587	0,0632
	MI	**0,0740**	0,0679	0,0659	0,0629	0,0612	0,0605	0,0604	0,0596	0,0591	0,0583	0,0630
Janela 20	LRD	0,1339	0,1361	0,1383	0,1388	0,1411	0,1412	0,1430	**0,1438**	0,1433	0,1421	0,1402
	Z Score	0,1129	0,1188	0,1191	0,1145	0,1151	0,1134	0,1175	0,1176	0,1180	**0,1210**	0,1168
	Phi-squared	**0,1038**	0,1033	0,1008	0,0965	0,0953	0,0924	0,0900	0,0889	0,0888	0,0870	0,0947
	VMI	**0,0765**	0,0683	0,0650	0,0630	0,0626	0,0615	0,0609	0,0595	0,0588	0,0586	0,0635
	MI	**0,0736**	0,0690	0,0656	0,0630	0,0626	0,0617	0,0611	0,0602	0,0589	0,0583	0,0634
Janela 50	LRD	0,1327	0,1334	0,1386	0,1402	0,1421	0,1413	0,1413	0,1411	0,1404	**0,1426**	0,1394
	Z Score	0,1179	0,1199	0,1179	0,1176	0,1209	0,1204	0,1239	0,1272	0,1269	**0,1295**	0,1222
	Phi-squared	**0,0881**	0,0866	0,0787	0,0811	0,0829	0,0790	0,0864	0,0880	0,0837	0,0796	0,0834
	MI	**0,0752**	0,0684	0,0664	0,0627	0,0619	0,0609	0,0598	0,0597	0,0595	0,0585	0,0633
	VMI	**0,0755**	0,0681	0,0644	0,0621	0,0628	0,0612	0,0605	0,0597	0,0589	0,0585	0,0632
Janela 100	LRD	0,1357	0,1338	0,1362	0,1375	0,1389	0,1406	0,1405	**0,1416**	**0,1416**	0,1403	0,1387
	Z Score	0,1197	0,1187	0,1182	0,1198	0,1211	0,1249	0,1208	0,1274	0,1281	**0,1295**	0,1228
	Phi-squared	**0,0968**	0,0878	0,0806	0,0749	0,0731	0,0731	0,0699	0,0662	0,0661	0,0664	0,0755
	VMI	**0,0753**	0,0675	0,0646	0,0619	0,0627	0,0616	0,0603	0,0595	0,0593	0,0586	0,0631
	MI	**0,0741**	0,0680	0,0657	0,0631	0,0612	0,0606	0,0605	0,0596	0,0591	0,0584	0,0630
Janela 200	LRD	0,1356	0,1337	0,1369	0,1374	0,1391	0,1400	0,1404	0,1411	**0,1428**	0,1405	0,1388
	Z Score	0,1199	0,1184	0,1190	0,1195	0,1203	0,1235	0,1212	0,1245	0,1260	**0,1249**	0,1217
	Phi-squared	**0,1096**	0,1020	0,0990	0,0905	0,0805	0,0808	0,0801	0,0791	0,0803	0,0786	0,0881
	VMI	**0,0753**	0,0676	0,0646	0,0620	0,1480	0,0617	0,0608	0,0594	0,0591	0,0586	0,0717
	MI	**0,0740**	0,0680	0,0660	0,0628	0,0611	0,0606	0,0604	0,0596	0,0591	0,0583	0,0630

Tabela 25 - Medida *F* média considerando dez fatores de expansão (*k*) e cinco configurações de janela

Nota: os maiores valores de *F* são destacados em negrito, e a ordem dos métodos em cada janela é definida pela média de todos os valores *k*

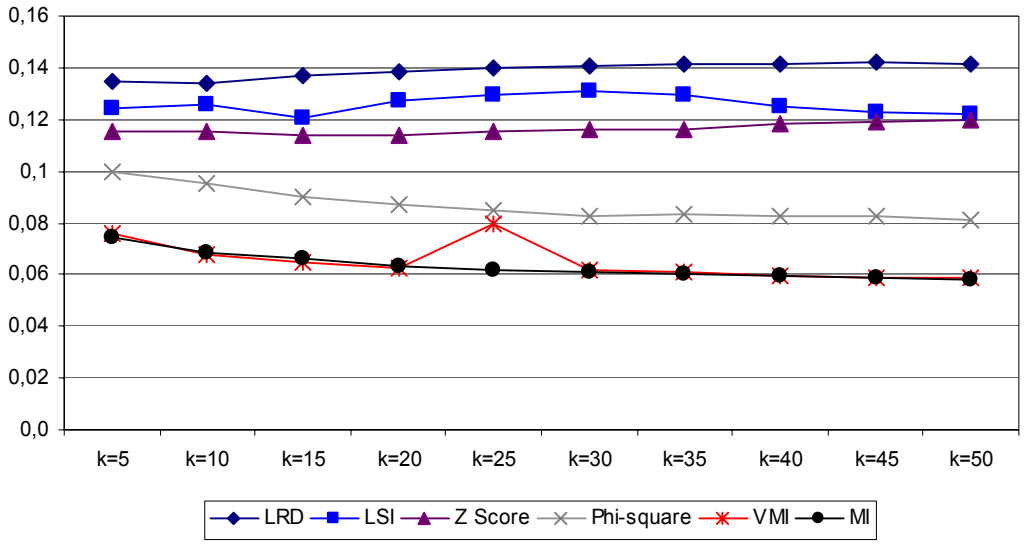


Figura 13 - Evolução da medida *F* média para os diversos fatores k considerando a média das cinco janelas

Tendo-se em vista que o índice-base é de 0,092, pode-se verificar que a expansão vetorial através dos métodos LRD, LSI e *Z score* tem efeito positivo sobre a recuperação de informação, enquanto que *Phi-squared,* MI e VMI possuem efeito negativo sobre a recuperação de informação. Levando-se em conta as simulações apresentadas na Tabela 25, verifica-se que LRD atinge o melhor desempenho para todas as janelas. Analisando-se o fator *k* em relação à média de todos os índices de expansão, sugere-se que valores de *k* entre 20 e 30, por estarem próximos da média, sejam adequados. Valores acima de *k*=30, além de não demonstrarem benefícios em relação ao desempenho da medida *F,* promovem incremento no custo computacional, visto que, quanto maior o *k*, maior o tempo de processamento.

Em termos de influência na variação do tamanho da janela, LRD, LSI e *Z score* praticamente não sofrem influência e consistentemente possuem melhor desempenho em relação aos demais métodos. De maneira similar, *Phi-squared*, MI e VMI apresentam pouca influência variando-se o tamanho da janela.

A Tabela 25 permite uma análise geral do desempenho do modelo de expansão vetorial aplicada à área de recuperação de informação. Entretanto, a média de todos os limiares não é adequada na utilização de aplicações reais. Analisando-se as informações contidas nos anexos do trabalho, verifica-se que, para o algoritmo LRD e mesmo para os demais métodos, o ponto de inflexão da medida *F* está em torno do limiar de 0.5 (ver Apêndice I).

A Tabela 26 apresenta os valores (entre 0 e 1) da medida *F* considerando o limiar de 0.5 e os cinco métodos de correlação, mais LSI para a configuração sem janela. Como mencionado anteriormente, LSI é o único método avaliado baseado simplesmente na utilização do documento completo em vez da utilização de janelas. Considerando-se LSI e os outros cinco métodos aplicados em diferentes configurações de janelas, LRD teve o melhor desempenho. O melhor índice da medida *F* é 0,1940 para janela de 20 e *k*=15. O segundo

método que melhor teve desempenho foi LSI, sendo a melhor medida *F* de 0,1720 para *k*=25 sem janela. O terceiro melhor desempenho foi obtido pelo método *Z score,* sendo a melhor medida *F* de 0,1735 para as janelas de 100 e *k*=40. Em seguida, está o *Phi-squared,* sendo o melhor resultado 0,1307 para a janela de 20 e *k*=10. Os métodos VMI e MI possuem, assim como na avaliação média, os piores desempenhos, mas são similares entre si.

Medida *F* (limiar 0,5)		*K*=5	*k=10*	*k*=15	*k=20*	*k*=25	*k=30*	*k*=35	*k=40*	*k*=45	*k=50*	*Média*
Sem Janela	**LRD**	0,1841	0,1808	0,1835	0,1872	0,1927	**0,1931**	0,1915	0,1886	0,1874	0,1782	0,1867
	LSI	0,1609	0,1565	0,1545	0,1614	**0,1720**	0,1664	0,1609	0,1521	0,1491	0,1490	0,1583
	Z Score	0,1277	**0,1295**	0,1219	0,1175	0,1172	0,1097	0,1074	0,1000	0,0950	0,0996	0,1126
	Phi-squared	0,1149	0,1084	0,1062	0,1083	0,1131	0,1087	0,1131	0,1130	0,1041	**0,1205**	0,1110
	VMI	**0,0663**	0,0611	0,0597	0,0581	0,0588	0,0577	0,0596	0,0593	0,0594	0,0591	0,0599
	MI	**0,0664**	0,0625	0,0601	0,0588	0,0579	0,0575	0,0572	0,0568	0,0574	0,0591	0,0594
Janela 20	**LRD**	0,1832	0,1871	**0,1940**	0,1892	0,1898	0,1904	0,1877	0,1874	0,1899	0,1905	0,1889
	Z Score	0,1431	0,1440	**0,1632**	0,1494	0,1492	0,1447	0,1540	0,1559	0,1568	0,1534	0,1514
	Phi-squared	0,1290	**0,1307**	0,1212	0,1069	0,1117	0,1199	0,1179	0,1171	0,1193	0,1113	0,1185
	VMI	**0,0671**	0,0617	0,0593	0,0589	0,0583	0,0579	0,0600	0,0596	0,0594	0,0590	0,0601
	MI	**0,0662**	0,0620	0,0594	0,0591	0,0586	0,0580	0,0579	0,0600	0,0595	0,0591	0,0600
Janela 50	**LRD**	0,1838	0,1816	0,1870	0,1881	**0,1892**	0,1891	0,1877	0,1893	0,1893	0,1870	0,1872
	Z Score	0,1451	0,1402	0,1520	0,1557	0,1541	0,1519	0,1562	0,1711	0,1590	**0,1611**	0,1546
	Phi-squared	**0,1063**	0,0970	0,0870	0,0955	0,1014	0,0934	0,1094	0,1075	0,1075	0,0993	0,1004
	MI	**0,0671**	0,0615	0,0609	0,0585	0,0585	0,0575	0,0572	0,0592	0,0597	0,0592	0,0599
	VMI	**0,0665**	0,0609	0,0596	0,0581	0,0589	0,0578	0,0575	0,0595	0,0594	0,0592	0,0597
Janela 100	**LRD**	0,1840	0,1811	0,1816	0,1864	0,1923	**0,1934**	0,1908	0,1882	0,1882	0,1771	0,1863
	Z Score	0,1463	0,1390	0,1511	0,1521	0,1501	0,1540	0,1534	**0,1735**	0,1604	0,1683	0,1548
	Phi-squared	**0,1037**	0,0791	0,0859	0,0808	0,0716	0,0709	0,0680	0,0656	0,0627	0,0627	0,0751
	VMI	**0,0663**	0,0610	0,0597	0,0581	0,0589	0,0579	0,0597	0,0593	0,0594	0,0591	0,0599
	MI	**0,0665**	0,0625	0,0600	0,0588	0,0580	0,0575	0,0572	0,0568	0,0574	0,0591	0,0594
Janela 200	**LRD**	0,1833	0,1814	0,1836	0,1872	0,1923	**0,1931**	0,1915	0,1880	0,1878	0,1783	0,1867
	Z Score	0,1458	0,1394	0,1510	0,1515	0,1491	0,1535	0,1535	0,1669	**0,1697**	0,1659	0,1546
	Phi-squared	**0,1303**	0,1203	0,0949	0,0870	0,0769	0,0771	0,0755	0,0755	0,0734	0,0715	0,0882
	VMI	**0,0663**	0,0611	0,0597	0,0581	0,1378	0,0577	0,0596	0,0593	0,0594	0,0591	0,0678
	MI	**0,0664**	0,0624	0,0601	0,0588	0,0580	0,0575	0,0572	0,0568	0,0574	0,0591	0,0594

Tabela 26 - Medida *F* considerando limiar de 0.5, dez fatores de expansão (*k*) e cinco configurações de janela

Nota: os maiores valores de *F* são destacados em negrito, e a ordem dos métodos em cada janela é definida pela média de todos os valores *k*

A Figura 14 apresenta a influência do fator *k* sobre os métodos utilizados. À proporção que *k* aumenta, a medida *F* do método LRD mantém-se praticamente estável. Para uma determinada configuração de janela, quando o fator *k* aumenta, a medida *F* dos métodos LRD e LSI aumenta até determinado *k* e depois tende a decrescer. Para a medida *F* do método LRD, o maior incremento acontece variando-se *k* de 20 para 25. Entretanto, conseqüentes

incrementos em *k* não demonstram ganhos significativos para o método. O método *Z score* apresenta oscilações, mas se mantém estável à medida que se varia *k*. Para os demais métodos à proporção que *k* varia, em geral, a medida *F* decresce. Tendo-se em vista que o índice-base é de 0,092, pode-se verificar que, na média de todos os índices *k*, a expansão vetorial através dos métodos LRD, LSI e *Z score* tem efeito positivo sobre a recuperação de informação, enquanto que *Phi-squared,* MI e VMI possuem efeito negativo. Levando-se em conta as simulações apresentadas na Tabela 26, verifica-se que LRD atinge o melhor desempenho para todas as janelas. Analisando-se o fator *k* em relação à média de todos os índices de expansão, sugere-se novamente que valores de *k* entre 20 e 30, por estarem próximos da média, sejam adequados. De maneira similar à análise anterior (Tabela 25), os valores acima de *k*=30, além de não demonstrarem benefícios em relação ao desempenho da medida *F,* promovem incremento no custo computacional, visto que incrementos em *k* interferem no tempo de processamento.

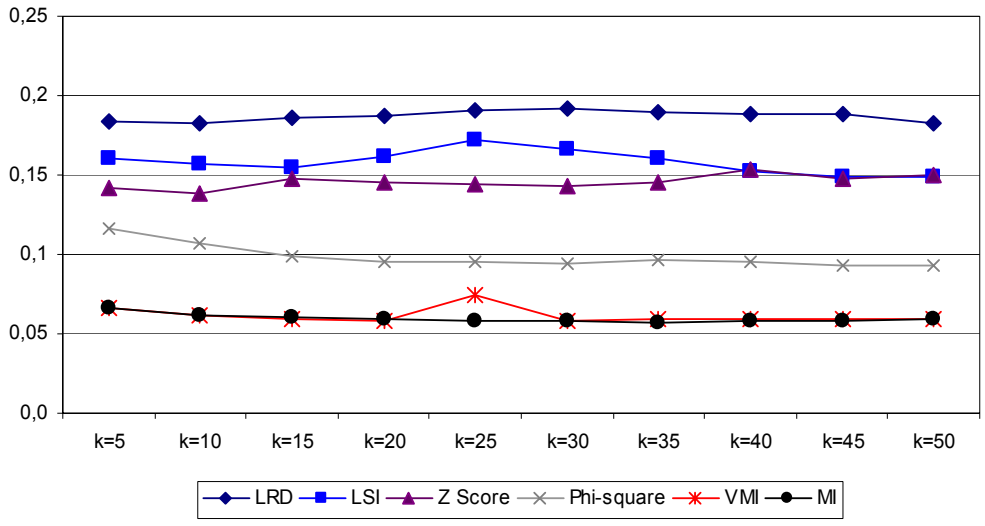


Figura 14 - Evolução da medida *F* (limiar de 0.5) para os diversos fatores *k* considerando a média das cinco janelas

Considerando-se a média dos fatores *k* em todas as janelas, verifica-se que LRD e *Z score* praticamente não sofrem influência e, juntamente com LSI, possuem melhor desempenho diante dos demais métodos. Por outro lado, *Phi-squared* e MI atingem,

considerando a média de todos os fatores *k*, o melhor resultado com a janela de 20, enquanto que VMI possui melhor resultado com janela igual a 200.

c) SUMARIZAÇÃO DOS RESULTADOS

A avaliação realizada sobre o conjunto de dados CISI utiliza-se da média das medidas *F* considerando limiares entre 0.1 e 0.8 bem como uma avaliação específica para o limiar 0.5. LRD é comparado com outros métodos possuindo, como ponto de referência a medida *F* de 0,092. Para o primeiro caso, ou seja, utilizando a média de todos os limiares, LRD obteve o melhor desempenho seguido pelos métodos LSI, Z score, Phi-squared, VMI e MI. Tendo-se em vista que o índice-base é de 0,092, verifica-se um efeito positivo na recuperação de informação para os métodos LRD, LSI e Z score, enquanto que Phi-squared, MI e VMI possuem efeito negativo sobre a recuperação de informação. Quanto ao fator *k*, sugerem-se valores entre 20 e 30 por estarem próximos da média. Valores acima de *k*=30, além de não demonstrarem benefícios em relação ao desempenho da medida *F*, promovem incremento no custo computacional, visto que, quanto maior o *k*, maior o tempo de processamento. O parâmetro de janela possui pouca influência tanto na média final quanto nos fatores *k* sugeridos (entre 20 e 30). Para a avaliação com o limiar de 0.5, a ordem de desempenho (LRD, LSI, Z score, Phi-squared, VMI e MI), o efeito positivo (LRD, LSI e Z score) e o efeito negativo (Phi-squared, MI e VMI) sobre a recuperação de informação não são alterados. Os valores dos fatores *k* são igualmente sugeridos entre 20 e 30 e o parâmetro de janela possui, similar à avaliação baseada na média, pouca influência.

4.1.2 AGRUPAMENTOS

Na avaliação do modelo de correlação e expansão vetorial utilizado na tarefa de agrupamento, o erro quadrático foi aplicado. Para tal, foi considerada a base REUTERS[20],

geralmente empregada na avaliação de aplicações de categorização de textos. Essa base possui um total de 19.043 documentos divididos em 45 classes. Através da análise do erro entre cada documento e a média do centróide ao qual o documento pertence, torna-se possível estabelecer o desempenho do método LRD proposto em relação a outras abordagens.

a) MODELO DE VALIDAÇÃO

Avaliar modelos não supervisionados constitui-se em uma tarefa não trivial. Uma das abordagens mais utilizadas é o somatório dos erros quadráticos. Seja n_i o número de exemplos no agrupamento C_i, a média desses exemplos m_i é definida como:

$$m_i = \frac{1}{n_i} \sum_{x \in C_i} x \,, \tag{24}$$

Assim o somatório dos erros quadráticos é definido como:

$$J = \sum_{i=1}^{C} \sum_{x \in C_i} \left\| x - m_i \right\|^2 \,, \tag{25}$$

Dessa forma, para um determinado agrupamento C_i, o vetor médio m_i é a melhor representação para os exemplos em C_i, visto que o somatório dos erros quadráticos é minimizado. Assim, J mede o erro quadrado total embutido na representação de n exemplos em c centros de agrupamentos. Essencialmente, o valor de J depende de como os exemplos são agrupados, sendo considerado um particionamento ótimo aquele que minimiza J.

O problema dessa abordagem reside na dependência do conhecimento prévio do número de agrupamentos (DUDA; HART 1973). Eventualmente isso pode ser obtido sobre pequenos conjuntos de dados ou quando padrões previamente determinados, possivelmente por intervenção humana, variam lentamente em função do tempo. Contudo, isso se torna impraticável, principalmente quando se está explorando um conjunto de dados desconhecido. O desafio é, portanto, definir a quantidade adequada de grupos em uma determinada análise.

Uma abordagem utilizada consiste em repetir o processo variando o número de agrupamentos *c* e analisar como o erro quadrático se modifica, uma vez que *J* deve decrescer monotonicamente[22] à medida que *c* aumenta. Se *n* exemplos em *c* grupos são compactos e bem separados, a tendência é que *J* decresça mais rapidamente até determinado *c*, e após isso decresça lentamente até atingir $c = n$. Nesse sentido, ao se atingir a fase de lenta redução do erro quadrático, o número de *c* para a análise em questão pode ser determinado por meio de alguma heurística.

b) DISCUSSÃO DOS RESULTADOS

Considerando-se os 10 valores de *k* utilizados na expansão vetorial em cinco diferentes configurações de janela, o total de conjuntos de dados para a tarefa de agrupamento é de 50. Cada conjunto de dados possui o vetor expandido para cada um dos métodos avaliados (LRD, *Phi-squared*, MI, VMI e *Z score*) nos respectivos fatores *k* e configurações de janela.

A expansão vetorial utilizou-se da base REUTERS composta de 19.043 documentos. Considerando-se cada fator *k* e as diferentes configurações de janela, a tarefa de agrupamento é realizada para cada um dos métodos avaliados utilizando-se a medida do co-seno. Para o limiar do co-seno (*n*) utilizado na formação dos agrupamentos, foram utilizados os valores de 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 e 0.8.

A equação do co-seno é empregada no processo de formação dos agrupamentos, ou seja, na determinação da similaridade dos vetores que formarão determinado agrupamento. Em cada processo de agrupamento, realizado sobre os vetores expandidos através de determinado método, fator *k,* configuração de janela e limiar *n*, o erro quadrático é determinado. De modo geral, à medida que *n* aumenta (variando de 0.1 até 0.8), o erro

[22] Uma função *f* é monotônica se sempre que $x \leq y$ então $f(x) \leq f(y)$, ou seja, uma função monotônica preserva a ordem. Nesse caso a função é chamada de monotonicamente crescente. Da mesma forma, uma função é chamada monotonicamente decrescente se sempre que $x \leq y$ então $f(x) \geq f(y)$. Neste caso *x* e *y* representam o número de agrupamentos.

quadrático também aumenta, uma vez que o número de agrupamentos reduz-se e a diferença de cada instância do conjunto de dados em relação à média de cada agrupamento aumenta.

O erro quadrático médio nos diferentes fatores *n* utilizando o espaço vetorial original, isto é, sem a expansão vetorial e sem janela, é de 15,7598[23]. Esse índice promove a base para a comparação com os demais métodos aplicados na expansão dos vetores. A Tabela 27 apresenta os erros médios, levando-se em conta todos os limiares (entre 0.1 e 0.8) para os cinco métodos de correlação analisados. Em todas as avaliações o método LRD teve o melhor desempenho. O menor erro ocorre para a configuração sem a janela. O segundo melhor desempenho foi obtido pelo método *Phi-squared* com janela igual a 100*,* seguido pelo método Z *score*, com janela de 200. Os métodos VMI e MI possuem o pior desempenho.

Erro		*k*=5	*k=10*	*K*=15	*k=20*	*K*=25	*k=30*	*k*=35	*k=40*	*k*=45	*k=50*	*Média*
Sem janela	**LRD**	10,9362	10,5116	10,0151	9,5096	8,9382	**8,2517**	**7,4779**	6,4373	4,7886	3,0271	7,9893
	Z Score	17,1293	15,8555	15,4661	15,2907	14,4924	**13,9433**	**13,4313**	12,0545	8,2014	9,4503	13,5315
	Phi-squared	16,8908	16,5894	15,9556	15,7470	15,2600	15,0919	**14,9647**	**14,5563**	9,6689	11,3873	14,6112
	VMI	16,9038	16,7857	16,7385	16,7331	16,7317	16,4421	16,1191	**15,8140**	**10,3268**	12,3151	15,4910
	MI	16,9038	16,7857	16,7385	16,7331	16,7317	16,4421	16,1191	**15,8140**	**10,2018**	12,3151	15,4785
Janela 20	**LRD**	16,7929	16,1097	15,3143	14,5098	13,6067	**12,4439**	**11,1170**	9,4084	6,9956	4,0606	12,0359
	Z Score	16,9214	16,5442	16,1362	15,6211	15,3191	**14,7465**	**13,8892**	12,9496	11,2442	8,9156	14,2287
	Phi-squared	16,3744	16,3111	16,2079	16,1937	16,1564	16,0558	15,7244	**15,5575**	**13,5517**	12,2618	15,4395
	VMI	17,3918	17,1703	16,8754	16,8607	16,5175	**16,0611**	**15,4720**	14,9673	13,8909	12,7424	15,7949
	MI	17,5660	17,5192	17,4849	17,4043	17,3058	17,0575	**16,7680**	**16,0839**	14,1502	12,2175	16,3557
Janela 50	**LRD**	13,5101	12,9499	12,3866	11,7032	10,9749	**10,1707**	**9,0471**	7,7772	5,6995	3,4339	9,7653
	Z Score	17,8562	17,3555	16,9368	16,4083	15,8898	**15,0195**	**14,1975**	13,1909	11,3883	8,2087	14,6451
	Phi-squared	16,6772	16,6114	16,6042	16,5507	16,3489	16,1860	**16,0209**	**15,6713**	13,7861	12,2800	15,6737
	VMI	18,1198	17,7099	17,6478	17,2787	17,2406	16,8833	**16,3243**	**15,5536**	14,1583	10,2553	16,1172
	MI	17,7793	17,6894	17,6350	17,6169	17,5141	17,0501	16,8457	**16,5754**	**14,4430**	12,3379	16,5487
Janela 100	**LRD**	12,0714	11,5340	11,0689	10,4696	9,8042	**9,0010**	**8,0530**	6,9459	5,1983	3,1903	8,7337
	Phi-squared	10,7143	10,5188	10,4025	10,2565	10,1462	**9,9615**	**9,8934**	9,6584	9,3558	7,6959	9,8603
	Z Score	16,7870	16,5980	16,1021	15,6877	15,2153	**14,5173**	**13,6202**	12,6070	10,6329	7,2424	13,9010
	VMI	16,8331	16,7203	16,7198	16,7141	16,6980	16,4057	**16,0887**	**15,7766**	13,8555	12,3067	15,8119
	MI	17,8440	17,7936	17,7135	17,7111	17,5776	17,1654	16,9071	**16,6625**	**14,5328**	12,3790	16,6287
Janela 200	**LRD**	11,2312	10,7495	10,2754	9,7665	9,1675	**8,4548**	**7,6034**	6,5658	4,9375	3,0594	8,1811
	Phi-squared	12,9856	12,7172	11,9947	11,6684	**11,3313**	**11,2323**	10,8904	10,7461	10,7134	8,4702	11,2750
	Z Score	15,1970	14,7409	14,2258	13,5032	13,1517	**12,4332**	**11,6775**	10,6957	8,8036	6,0402	12,0469
	VMI	16,8732	16,7534	16,7401	16,7338	16,7328	16,4353	**16,1186**	**15,8137**	13,8868	12,3156	15,8403
	MI	17,8906	17,8283	17,7557	17,7488	17,5825	17,1461	16,9255	**16,7010**	**14,5504**	12,3812	16,6510

Tabela 27 - Erro quadrático considerando dez fatores de expansão (*k*) e cinco configurações de janela

Nota: a ordem dos métodos em cada janela é definida pela média de todos os valores *k*

[23] Valor dividido por 1000.

Analisando-se a média de todos os métodos nas diferentes janelas, é possível determinar um provável fator de expansão (*k*) entre 30 e 40. De modo geral, à medida que *k* aumenta (variando de 5 até 50), o erro quadrático diminui, uma vez que cada instância do conjunto de dados (vetor) e os próprios agrupamentos se tornam mais densos. Isso reduz a diferença dos vetores em relação aos centros dos agrupamentos e conseqüentemente o erro quadrático. Tendo-se em vista que o erro-base médio é de 15,7598, pode-se verificar que a expansão vetorial através dos métodos LRD, *Phi-squared* e *Z score* tem efeito positivo sobre a tarefa de agrupamento, enquanto que VMI e MI possuem na maioria das vezes efeito negativo sobre a tarefa de agrupamento.

Em termos de influência na variação do tamanho da janela, LRD obtém resultados melhores sem qualquer restrição de janela, ou seja, à medida que a janela aumenta, o erro quadrático diminui. Essa é uma característica importante, uma vez que não se torna necessário especificar tal parâmetro para que o método atinja os resultados adequados. Basicamente, LRD consistentemente melhora o contexto de representação de documentos à medida que o parâmetro da janela é incrementado.

O estabelecimento *a priori* da média de um conjunto de limiares em aplicações reais não é desejável devido ao tempo computacional. Para uma análise individualizada e considerando-se a média de todos os limiares, o erro mais próximo da média fica entre os limiares de 0.5 e 0.6 (ver Apêndice II). No presente trabalho será utilizado o limiar de 0.5.

A Tabela 28 apresenta os erros com base no limiar de 0.5 para os cinco métodos de correlação analisados. Em todas as avaliações o método LRD teve o melhor desempenho. O menor erro ocorre para a configuração sem a janela. O segundo melhor desempenho foi obtido pelo método *Phi-squared*, com uma janela de 100, seguido pelo método *Z score* sem restrição de janela. Os métodos VMI e MI possuem o pior desempenho, mas abaixo do erro de 15,7598, obtido utilizando-se os vetores originais.

Analisando-se a média de todos os métodos nas diferentes janelas, é possível determinar um provável fator de expansão (*k*) entre 30 e 40. Considerando-se que o erro-base é de 15,7598, pode-se verificar que a expansão vetorial através dos métodos avaliados e utilizando-se limiar de 0.5 possui efeito positivo em todas as janelas.

	Erro	*k*=5	*k=10*	*K*=15	*k=20*	*k*=25	*k=30*	*k*=35	*k=40*	*k*=45	*k=50*	*Média*
Sem janela	**LRD**	7,5305	7,2296	6,9122	6,5935	6,2140	5,7844	**5,2168**	**4,5096**	3,3440	1,8650	5,5200
	Z Score	10,0869	9,6651	9,3716	9,1501	**8,8052**	**8,3087**	8,0755	7,4945	6,7592	5,8063	8,3523
	Phi-squared	11,3105	11,1639	10,9802	10,6272	10,4085	10,2579	**10,2311**	**9,9891**	9,3506	7,8226	10,2141
	VMI	12,1800	12,0950	12,0610	12,0570	12,0561	11,8473	**11,6169**	**11,3987**	9,9943	8,7898	11,4096
	MI	12,1800	12,0950	12,0610	12,0570	12,0561	11,8473	**11,6169**	**11,3987**	9,9943	8,7898	11,4096
Janela 20	**LRD**	11,6061	11,1370	10,5890	10,0591	9,4351	**8,6801**	**7,7439**	6,5527	4,9090	2,6472	8,3359
	Z Score	11,9289	11,6865	11,3889	11,0459	10,7615	**10,3544**	**9,7263**	9,0717	7,7980	6,1977	9,9960
	Phi-squared	11,7984	11,7528	11,6797	11,6682	11,6413	11,5689	11,3325	**11,2038**	**9,7686**	8,7394	11,1154
	VMI	12,2915	12,2169	12,0320	11,9763	11,7021	**11,4091**	**11,0018**	10,6334	9,8610	9,0369	11,2161
	MI	12,6570	12,6233	12,5985	12,5404	12,4697	12,2906	**12,0835**	**11,5866**	10,1721	8,7479	11,7770
Janela 50	**LRD**	9,2809	8,8923	8,5102	8,0508	7,5926	**7,0444**	**6,3023**	5,3795	4,0062	2,1889	6,7248
	Z Score	12,6305	12,3130	12,0068	11,6178	11,2714	**10,6471**	**10,0849**	9,3448	8,0716	5,7997	10,3788
	Phi-squared	12,0168	11,9695	11,9642	11,9257	11,7802	11,6628	11,5456	**11,2950**	**9,9260**	8,7533	11,2839
	VMI	12,9103	12,6269	12,5748	12,3973	12,2139	11,9767	**11,5236**	**11,0525**	10,1662	7,2010	11,4643
	MI	12,8108	12,7461	12,7069	12,6940	12,6197	12,2855	12,1406	**11,9479**	**10,3898**	8,8085	11,9150
Janela 100	**LRD**	8,2817	7,9387	7,5981	7,2062	6,7844	**6,2892**	**5,6692**	4,8473	3,6144	1,9816	6,0211
	Phi-squared	7,7245	7,5588	7,5221	7,4319	7,2967	7,1551	**7,1480**	**6,9497**	6,7838	5,4166	7,0987
	Z Score	11,9078	11,7426	11,3824	11,1145	10,7497	**10,2898**	**9,6442**	8,9333	7,5227	5,1811	9,8468
	VMI	12,1291	12,0479	12,0472	12,0434	12,0318	11,8211	**11,5950**	**11,3711**	9,9757	8,7733	11,3836
	MI	12,8576	12,8213	12,7636	12,7619	12,6655	12,3686	12,1850	**12,0109**	**10,4538**	8,8335	11,9722
Janela 200	**LRD**	7,7175	7,4111	7,0878	6,7482	6,3717	**5,9127**	**5,3357**	4,6062	3,4134	1,8922	5,6496
	Phi-squared	9,3771	9,1822	8,6575	8,4407	**8,1760**	**8,0176**	7,8558	7,7490	7,7246	5,6904	8,0871
	Z Score	10,8203	10,4957	10,1034	9,5829	**9,3192**	**8,8043**	8,2887	7,5743	6,2430	4,3178	8,5550
	VMI	12,1580	12,0717	12,0621	12,0576	12,0569	11,8424	**11,6166**	**11,3985**	10,0054	8,7889	11,4058
	MI	12,8912	12,8463	12,7941	12,7891	12,6691	12,3547	12,1969	**12,0369**	**10,4650**	8,8373	11,9881

Tabela 28 - Erro quadrático considerando limiar de 0.5, dez fatores de expansão (*k*) e cinco configurações de janela

Nota: a ordem dos métodos em cada janela é definida pela média de todos os valores *k*

Em termos de influência na variação do tamanho da janela, LRD obtém melhores resultados sem qualquer restrição de janela, ou seja, à medida que a janela aumenta o erro quadrático diminui. Similar à análise anterior, LRD consistentemente melhora o contexto de representação de documentos à medida que o parâmetro da janela é incrementado.

c) SUMARIZAÇÃO DOS RESULTADOS

A avaliação realizada sobre o conjunto de dados REUTERS baseia-se no erro quadrático médio considerando limiares entre 0.1 e 0.8 e uma avaliação específica para o limiar 0.5. LRD é comparado com outros métodos, possuindo como ponto de referência um erro quadrático médio de 15,7598. Para o primeiro caso, ou seja, utilizando a média, LRD obteve o melhor desempenho, seguido pelo método *Phi-squared*, *Z score*, VMI e MI. Tendo-se em vista que o índice-base é de 15,7598, verifica-se um efeito positivo na tarefa de agrupamento para os métodos LRD, *Phi-squared*, *Z score*, enquanto que MI e VMI possuem efeito negativo nessa tarefa. Quanto ao fator *k* sugerem-se valores entre 30 e 40 por estarem próximos da média. Considerando-se o parâmetro de janela, LRD obteve resultados melhores sem qualquer restrição, ou seja, à medida que a janela aumenta, o erro quadrático diminui. Para os demais métodos esse parâmetro possui pouca influência, não sendo possível determinar qual a melhor configuração. Para a avaliação com o limiar de 0.5, todos os métodos possuem efeito positivo na tarefa de agrupamento. Os valores dos fatores *k* são igualmente sugeridos entre 30 e 40. O parâmetro de janela possui pouca influência uma vez que em geral os melhores resultados foram obtidos sem qualquer restrição na janela.

4.1.3 VALIDAÇÃO ORIENTADA À TAREFA

Esta abordagem é útil para uma análise inicial do potencial de determinado método, mas, como mencionado anteriormente, seu projeto deve ser cuidadoso, visando minimizar erros introduzidos através da subjetividade. Nesse sentido, dois estudos são realizados, sendo o primeiro relativamente mais simples e com um grau maior de subjetividade na análise, pois sugere a ordem de relevância das entidades, e um segundo mais abrangente, em que somente se questiona a relevância de determinada relação entre as entidades sem qualquer ordem. Esses dois estudos são demonstrados nos casos A e B, descritos a seguir.

a) CASO A

Neste primeiro estudo foi considerado um *Website* departamental com 503 páginas, sendo extraídas entidades de quatro classes, Pessoa, Organização, Projetos e Áreas de Pesquisa. O estudo possui como objetivo comparar os índices de precisão, lembrança e acurácia de ordenação para o modelo LRD no estabelecimento da força do relacionamento entre os membros do departamento (entidades do tipo pessoa) e suas entidades relacionadas (ERs) nas quatro classes. De modo a definir o limiar de relevância, ou seja, o ponto de corte entre as ERs consideradas corretas e incorretas, três pessoas foram selecionadas para a avaliação. Cada pessoa analisou 10 páginas em que seu nome se relacionava com outras entidades, verificando assim que, ao estabelecer o limiar de distância de 10, 92% das ERs dentro do limiar foram julgadas relacionadas pelas três pessoas. O limiar é calculado através da Equação 18, modificando-se a função de distância para $\bar{d} = 1 + \log_2(n)$, onde n=10. Considerou-se ainda que, para uma relação ser válida, ela deve co-ocorrer pelo menos uma vez em pelo menos uma página. O valor final do limiar é de $R(E1, E2) = \frac{1}{503} x \frac{1}{1 + \log_2(10)} = 4.6x10^{-4}$.

Para a avaliação foram selecionadas 20 pessoas de um universo de 60, representando estudantes, pesquisadores e professores. Pediu-se de maneira independente que fosse realizada uma análise das relações pessoais através de um formulário que apresentava listas de ERs, de acordo com a classe e por ordem de relevância (Figura 15).

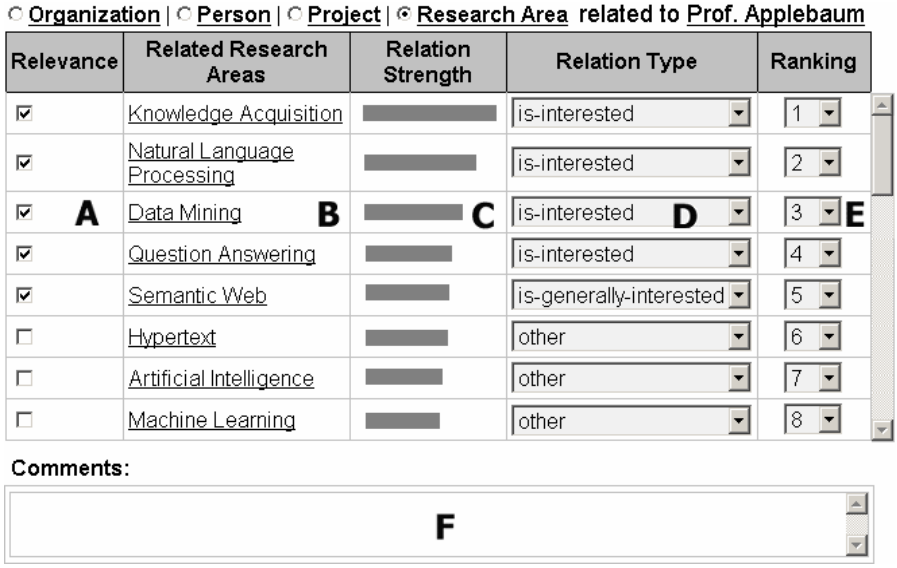


Figura 15 - Exemplo de formulário de avaliação das entidades relacionadas

Herlocker et al. (2004) fazem uma revisão sobre Sistemas de Recomendação Colaborativa (SRC). Essa classe de sistemas utiliza-se da opinião de uma comunidade de usuários para a realização de sugestões. A precisão e a lembrança, duas das mais populares métricas para avaliação de sistemas de recuperação de informação, são utilizadas nesse cenário para medir a habilidade de sistemas RC indicarem itens relevantes. A medida *F* combina essas duas métricas. Adicionalmente, métricas de correlação são utilizadas para avaliar a capacidade de sistemas RC sugerirem listas ordenadas de itens que sigam uma lista ideal, sugerida pelo usuário. Similarmente, essas métricas são utilizadas na avaliação do modelo LRD ao prover relevantes listas de ERs a partir de uma entidade de origem.

Para uma lista de ERs de tipo *T*, o número de ERs julgadas relevantes por LRD é $N_{LRD,\mathrm{Relevante}}$. Após a avaliação do usuário, o número total de ERs julgadas como relevantes pelo usuário é $N_{Usuário,\mathrm{Relevante}}$. O número total de ERs julgadas como relevantes por ambos – usuário e LRD – é definido por $N_{Usuáro,LRD,\mathrm{Relevante}}$. A precisão $P_{T,Usuário}$ e a lembrança $R_{T,Usuário}$ são definidas nas equações 26 e 27.

$$P_{T,Usuário} = \frac{N_{Usuário,LRD,\text{Relevante}}}{N_{LRD,\text{Relevante}}} \quad (26)$$

$$R_{T,Usuário} = \frac{N_{Usuário,LRD,\text{Relevante}}}{N_{Usuário,\text{Relevante}}} \quad (27)$$

LRD e usuário provêem duas listas de ERs ordenadas $N_{Usuário,LRD,\text{Relevante}}$ julgadas relevantes por ambos. Para medir o grau de combinação entre LRD e usuário, $RA_{T,Usuário}$ é definido como o coeficiente de correlação de Spearman (GIBBONS, 1976) entre dois conjuntos ordenados, como apresentado na Equação 28.

$$RA_{T,Usuário} = 1 - \frac{6\sum_{i}(r_{i,Usuário} - r_{i,LRD})^2}{N_{Usuário,LRD,\text{Relevante}}{}^3 - N_{Usuário,LRD,\text{Relevante}}}, \quad (28)$$

onde $r_{i,Usuário}$ e $r_{i,LRD}$ são duas listas ordenadas providas pelo usuário e LRD, respectivamente, para a *i*th ER na lista. O valor resultante é $-1 \le RA \le 1$ onde *RA*=1 quando duas listas estão em perfeito acordo e *RA*=-1 quando não existe nenhum acordo.

As Equações 26, 27 e 28 são utilizadas para calcular os índices de precisão, lembrança e correlação de quatro listas de ERs, sendo *OrgRES*, *PesRES*, *ProjRES* e *AreaRES*, para cada um dos 20 usuários. Na Figura 16 observa-se que os índices de precisão e lembrança para todas as REs e todos os 20 usuários ficam entre 70% e 100%, e o índice de correlação, entre 0 e 1.0.

O número total de ERs para os quatro tipos e todos os usuários julgados relevantes por LRD é $Total_{LRD,\text{Relevante}}$, o número total de ERs julgadas relevantes pelos usuários é $Total_{Usuário,\text{Relevante}}$, e o número total de ERs julgadas relevantes pelos usuários e LRD é $Total_{LRD,Usuário,\text{Relevante}}$. As equações globais de precisão (P_{Total}), lembrança (R_{Total}) e medida *F* (F_{Total}) usadas nas avaliações são definidas como:

$$P_{Total} = \frac{Total_{LRD,Usuário,\text{Relevante}}}{Total_{LRD,\text{Relevante}}} \quad (29)$$

$$R_{Total} = \frac{Total_{LRD,Usuário,\text{Relevante}}}{Total_{Usuário,\text{Relevante}}} \quad (30)$$

$$F_{Total} = \frac{2R_{Total} \times P_{Total}}{R_{Total} + P_{Total}} \tag{31}$$

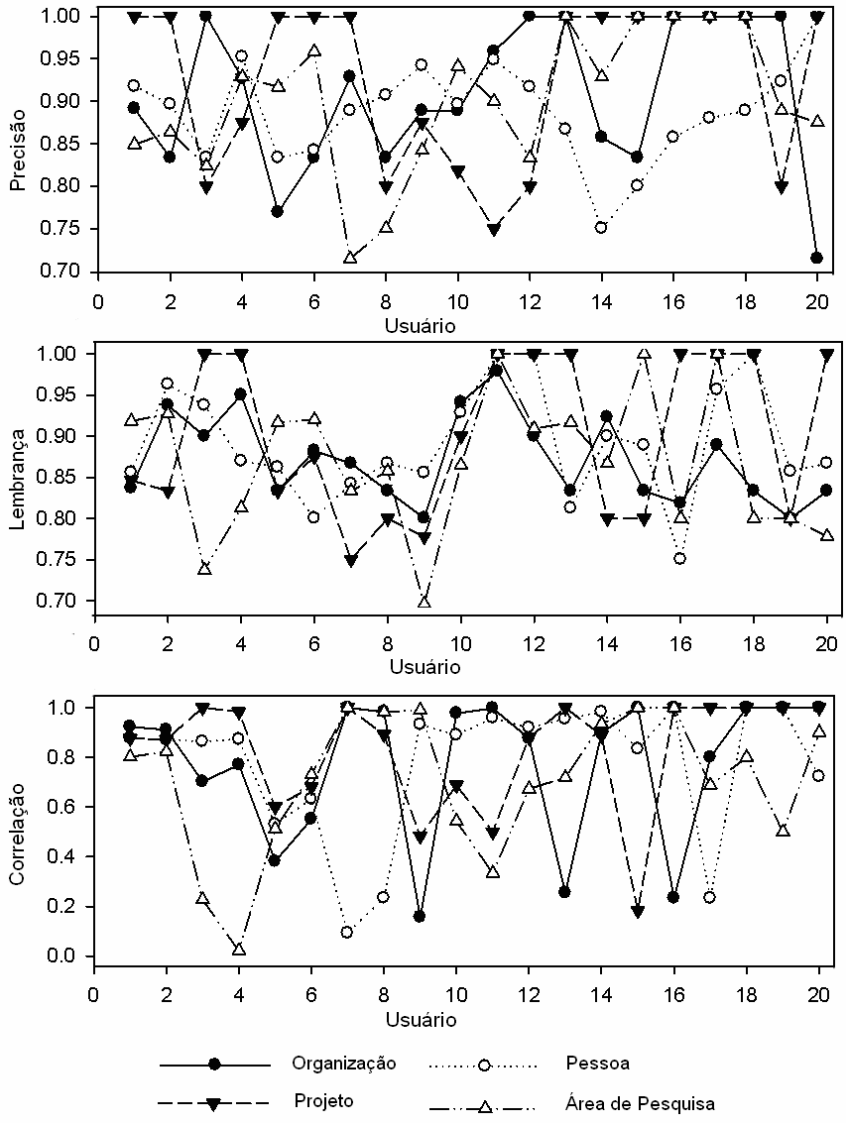


Figura 16 - Índices de precisão, lembrança e correlação da avaliação

Os valores obtidos na avaliação referentes a P_{Total}, R_{Total} e F_{Total} são de 0,905, 0,882 e 0.904, respectivamente. Diferentes limiares para a força do relacionamento foram utilizados visando ajustar P_{Total} em relação à R_{Total}, e o melhor resultado foi atingido utilizando-se 4.6×10^{-4}. Para a acurácia da ordenação, calculou-se a média de todos os 20 usuários, e o valor final atingido para RA_{Total} é de 0,769.

De maneira geral, todos os 20 usuários classificaram os resultados produzidos pelo método LRD como úteis ou muito úteis. Entretanto, determinadas considerações foram

efetuadas. Sugeriu-se que ERs altamente relacionadas entre si, por exemplo, "mineração de dados" e "mineração de textos" para a coleção de documentos em análise, fossem agrupadas para facilitar o julgamento. Os usuários tiveram dificuldade em ordenar certos tipos de ERs, tais como *PesREs*, uma vez que suas percepções pessoais dos níveis de importância das ERs são difíceis de serem quantificadas. Para tal, sugeriu-se que as relações pudessem considerar a questão temporal, visto que relações mais atuais tendem a ser mais relevantes.

Com relação às ERs julgadas como relevantes pelo método LRD, mas irrelevantes pelos usuários, verificou-se que a maioria estava um pouco acima do limiar estabelecido. Em muitos casos isso aconteceu justamente pela falta do componente temporal na equação. Para ERs julgadas irrelevantes pelo método LRD, mas relevantes pelos usuários, verificou-se que a maioria estava ligeiramente abaixo do limiar. Isso se justifica pela falta de informação sobre as relações no contexto considerado, ou seja, na coleção de documentos utilizada. Com relação à avaliação da ordenação das ERs pelos usuários, na maioria dos casos somente se realizaram mudanças regionais em vez de mudanças globais, ou seja, acima ou abaixo do limiar. Apesar de várias limitações, tais como erros ou falta de informação na coleção de documentos, erros no processo de extração de entidades e subjetividade da avaliação provida pelos usuários, o método produziu resultados satisfatórios.

a.1) SUMARIZAÇÃO DOS RESULTADOS

A avaliação foi realizada utilizando um *Website* departamental como fonte de informação assim como o julgamento de 20 especialistas para analisar a relevância das relações entre entidades sugeridas pelo método LRD. Para a mensuração dos resultados, aplicaram-se as medidas de precisão, lembrança, medida *F* e índice de correlação. Os valores obtidos na avaliação referentes a P_{Total}, R_{Total}, F_{Total} e RA_{Total} são de 0,905, 0,882, 0,904, e 0,769, respectivamente. Diferentes limiares para a força do relacionamento foram utilizados visando

ajustar P_{Total} em relação à R_{Total}, e o melhor resultado foi atingido utilizando-se 4.6×10^{-4}. De maneira geral os usuários identificaram os resultados produzidos pelo método como relevantes. Apesar das diferenças de julgamento entre o método e os usuários, notou-se que as mudanças, quando ocorreram, foram realizadas regionalmente em vez de globais, ou seja, acima ou abaixo do limiar, mas não entre as regiões.

b) CASO B

Este estudo vislumbra a possibilidade de localização de especialistas em determinado assunto através da avaliação de seus pares dentro de uma organização. Desse modo, assume-se, assim como um sistema de votação, que especialistas possam ser identificados por meio de um sistema de coleta de opiniões. Tais opiniões são então utilizadas como ponto de equilíbrio na avaliação de cinco abordagens de ordenação de listas de preferências. Objetiva-se comparar o grau de concordância das sugestões de especialistas em determinado assunto efetuadas pelos participantes do estudo e os métodos de ordenação/correlação.

Para tal, foram utilizadas informações de dois *Websites* departamentais contendo 1.100 e 1.900 páginas, respectivamente, sendo consideradas as classes de entidades “Pessoas” e “Áreas de Pesquisa”. Dois grupos de pessoas, um de 23 e outro de 17, respectivamente, foram selecionados para os dois departamentos. Os grupos são formados por estudantes, desenvolvedores e doutores. Cada pessoa avaliou um total de 13 consultas. Para cada consulta os modelos LRD, *Phi-squared*, MI, VMI e *Z score* foram aplicados, considerando-se o valor da janela de 100.

De modo geral, um modelo de ordenação pode produzir longas listas de especialistas, contudo, assume-se que os usuários somente mantêm interesse nos primeiros itens. Nesse sentido somente os 10 primeiros especialistas (esse número pode ser menor) retornados em cada modelo para cada consulta são levados em conta. Para determinada consulta, cada um

dos modelos pode não somente ordenar especialistas diferentemente como também produzir diferentes listas de especialistas, que agrupadas formam a lista final para a consulta em questão, a qual geralmente supera 10 itens.

Para a avaliação elaborou-se um formulário que possibilita o julgamento dos resultados gerados pelos modelos de correlação (Figura 17). De modo a minimizar a influência da ordem dos especialistas em cada consulta, a lista é apresentada de maneira aleatória, sem qualquer referência à relevância da relação. Para cada consulta pediu-se que o usuário informasse o seu grau de concordância entre a especialidade (consulta) e o especialista, sendo as opções disponíveis “discorda fortemente”, -2, “discorda”, -1, “concorda em parte”, 0, “concorda”, 1, ou “concorda fortemente”, 2. Dessa forma, para cada consulta e especialista relacionados, é calculada a média, possibilitando o estabelecimento de um modelo *post hoc* utilizado na avaliação do desempenho dos modelos de correlação.

Consulta = "Machine learning"

Nome	Sem opinião	Discorda Fortemente	Discorda	Concorda em parte	Concorda	Concorda Fortemente
Especialista 1	◉	○	○	○	○	○
Especialista 2	◉	○	○	○	○	○
Especialista 3	◉	○	○	○	○	○
Especialista 4	◉	○	○	○	○	○
Especialista 5	◉	○	○	○	○	○
Especialista 6	◉	○	○	○	○	○

Figura 17 - Exemplo de formulário de avaliação de especialistas relacionados à consulta

O grau de concordância entre as respostas fornecidas pelos usuários e os modelos de correlação é determinado pelo coeficiente de Spearman (Equação 32).

$$RA_{Método,Grupo} = 1 - \frac{6\sum_{i}(r_{i,Grupo} - r_{i,Método})^2}{N^3 - N} \tag{32}$$

onde $r_{i,Grupo}$ e $r_{i,Método}$ são duas listas ordenadas providas pelo grupo de usuários que avaliou determinada consulta e pelos métodos de correlação referindo-se ao *i*th especialista na lista. O

valor resultante é $-1 \leq RA \leq 1$ onde *RA*=1 quando duas listas estão em perfeito acordo e *RA*=-1 quando não existe nenhum acordo.

A Tabela 29 e a Tabela 30 apresentam os níveis de concordância entre os dois grupos de avaliadores e os modelos de correlação utilizados. Analisando-se a Tabela 29, verifica-se que LRD, MI, Phi-squared, VMI e *Z score* obtiveram os melhores resultados nas consultas 11, 7, 11, 9, 11, respectivamente, ao passo que, analisando-se a Tabela 30 o melhor resultado obtido pelo método LRD foi na consulta 12. Todas as demais medidas tiveram o melhor desempenho na consulta 9. LRD obteve melhor desempenho na maioria das consultas. Todavia, na média, seu desempenho foi similar aos métodos *Phi-squared* e *Z score*. Os valores obtidos neste estudo pelo método LRD de 0,7114 e 0,5251, nos dois grupos, são similares ao estudo anterior, que obteve um índice de 0,769. Essa redução pode ser explicada em parte pelo aprimoramento do estudo (projeto), que, em vez de propor a lista de ERs previamente ordenada, apresenta-a de maneira aleatória para o julgamento.

Consulta/Relevância de Ordenação	LRD	MI	Phi-squared	VMI	Z score
1. Artificial intelligence	**0,9140**	-0,0351	0,8061	-0,1044	0,7781
2. Hypertext	0,6838	0,3320	0,6093	0,1993	**0,7007**
3. Information extraction	0,7047	0,0735	**0,8137**	-0,1397	0,4730
4. Knowledge management	**0,9200**	0,1863	0,8229	0,1244	0,7101
5. Knowledge modelling	**0,9596**	-0,0114	0,7667	-0,2298	0,7702
6. Machine learning	**0,5839**	0,0559	0,2378	0,0909	0,1329
7. Natural language processing	-0,1545	**0,5636**	0,3182	0,5000	0,3455
8. Ontologies	**0,9466**	0,4628	0,9313	0,1086	0,7337
9. Planning AND Scheduling	**0,7657**	0,5000	0,6538	0,6189	0,6818
10. Question answering	0,6171	0,0939	0,6089	0,1228	**0,6264**
11. Semantic web	0,9688	0,1719	**0,9846**	0,3169	0,9285
12. Web services	0,6054	0,4632	**0,6569**	0,3333	0,6397
13. Social software	**0,7335**	0,2487	0,5552	0,2591	0,3361
Média	**0,7114**	0,2389	0,6743	0,1693	0,6044

Tabela 29 - Nível de concordância para o primeiro grupo

Consulta/Relevância de Ordenação	LRD	MI	Phi-squared	VMI	Z score
1. Artificial intelligence	**0,7821**	-0,3107	0,1607	-0,2286	0,1661
2. Europe AND learning	**0,5797**	0,2967	0,4863	0,2637	0,5027
3. Human learning	0,0909	0,2091	0,2818	**0,3727**	**0,3273**
4. Hypermedia	**0,5265**	0,3000	0,4147	0,3368	**0,5265**
5. Hypertext	**0,5385**	-0,0412	0,3159	0,0742	0,4231
6. Instant messaging	0,4643	0,3929	**0,5714**	0,3929	**0,5714**
7. Java AND C++	**0,5515**	0,1273	0,2485	0,1273	0,1636
8. Knowledge representation	**0,6905**	0,3095	0,3095	0,3095	0,3095
9. Learning AND grid	**0,6835**	0,4923	0,6091	0,6099	0,6808
10. Machine learning	**0,4286**	0,1429	0,1429	0,1786	0,1786
11. Semantic web	0,2445	-0,1291	**0,3104**	-0,2335	0,2060
12. Social software	**1,0000**	-1,0000	0,5000	-1,0000	0,5000
13. Software AND grid	0,2451	0,1483	0,3358	0,3370	**0,3909**
Média	**0,5251**	0,0722	0,3605	0,1185	0,3805

Tabela 30 - Nível de concordância para o segundo grupo

De modo geral, LRD, *Phi-squared* e *Z score* produzem associações de entidades com alta freqüência, enquanto MI e VMI favorecem associações de baixa freqüência. Considerando-se que associações com alta freqüência geralmente indicam uma maior importância em determinada relação, LRD, *Phi-squared* e *Z score* são mais adequados para esse tipo de tarefa.

Verifica-se ainda que, quando a concordância do método LRD é baixa, isso geralmente indica pouca informação para determinada consulta. Um exemplo disso é a consulta "*Natural Language Processing*" apresentada na Tabela 29. Esse tópico não faz parte das principais áreas de pesquisa do departamento em questão (proporcionalmente existem poucas páginas mencionando qualquer relação com o tópico), e muitos avaliadores tinham pouca clareza sobre quem estaria associado a essa consulta.

b.1) SUMARIZAÇÃO DOS RESULTADOS

A avaliação foi realizada utilizando dois *Websites* departamentais como fonte de informação bem como o julgamento de 40 pessoas. Possui como objetivo medir a acurácia (grau de concordância) das sugestões de especialistas efetuadas para uma determinada

especialidade através dos métodos LRD, *Phi-squared*, MI, VMI e *Z score*. Cada pessoa avaliou um total de 13 consultas, que requisitavam ao usuário que informasse o seu grau de concordância entre a especialidade (consulta) e o especialista, sendo as opções disponíveis "discorda fortemente", -2, "discorda", -1, "concorda em parte", 0, "concorda", 1, ou "concorda fortemente", 2. O melhor resultado, nos dois conjuntos de dados, foi obtido pelo método LRD com 0,7114 e 0,5251, respectivamente, seguido pelos métodos *Phi-squared* e *Z score*.

4.1.4 VALIDAÇÃO GERAL DO MÉTODO LRD

Com o objetivo de realizar um teste geral do modelo LRD, será utilizada uma base de entidades correlacionadas. Essa base foi gerada a partir de um conjunto de 300 artigos na área de Web Semântica. Para cada artigo, é extraído o vetor de entidades contendo as informações necessárias ao cálculo. As entidades estão distribuídas em três classes, sendo 970 organizações, 914 pessoas e 417 áreas de pesquisa, totalizando 2.301 entidades. Procurando-se evitar análises subjetivas (abordagem orientada a tarefas), as relações entre uma entidade origem e suas entidades relacionadas são avaliadas, confrontando-se a freqüência conjunta de um determinado mecanismo de busca e a força do relacionamento calculado pelos métodos de correlação analisados no trabalho.

Nessa abordagem argumenta-se que a freqüência conjunta, obtida a partir do número total de documentos que mencionam o par de entidades $r(E1, E2)$, é um indício de que existe relacionamento entre elas. Nesse sentido, seja uma entidade $E1$, um mecanismo de busca foi utilizado para determinar a freqüência conjunta com as entidades relacionadas, a qual é também um indicativo da ordem em que os pares relacionados estão dispostos. Essa lista serve de base para que os métodos de correlação de elementos textuais sejam comparados. Para se avaliar a precisão da correlação ou ordenação, aplica-se a equação de *Spearman*:

$$RA = 1 - \frac{6\sum_{i}(R_{i,MC} - R_{i,MB})^2}{N^3 - N}, \quad (33)$$

onde -1 ≤ RA ≤ 1, $r_{\mathrm{i,MC}}$ refere-se à ordem do item para um específico método de correlação, $r_{\mathrm{i},MB}$, à ordem obtida através do mecanismo de busca, e *N,* ao número de entidades recuperadas para a consulta.

No processo de avaliação todas as 2.301 entidades foram consideradas. Para cada entidade foram selecionadas as *n* entidades mais relevantes em cada um dos modelos de correlação considerados, LRD (*M*1), *Phi-squared* (*M*2), MI (*M*3), VMI (*M*4) e *Z score* (*M*5). Para o estudo em questão, *n* possui o valor de 10. Diferentes entidades podem ser selecionadas pelos métodos, o que ao final, geralmente, irá produzir uma lista de relações superior a 10 pares. Para cada entidade utilizada na consulta foram consideradas diferentes configurações de janelas, ou seja, 20, 50, 100, 200 e sem janela.

A Tabela 31 apresenta um exemplo utilizando o termo "*Semantic Web*" (entidade do tipo área de pesquisa) para uma janela de 20 termos. Pode-se observar que os melhores resultados foram obtidos pelo modelo LRD, *Phi-squared* e *Z score*, com 0.930, 0.601 e 0.372, respectivamente.

Se uma determinada entidade não é selecionada pelo método, o índice atribuído de *M*1 até *M*5 será *n*+1. Se a posição de uma entidade, de *M*1 até *M*5, for igual a *n*+1 e o índice MB for menor do que *n*+1 ou a posição de uma entidade de *M*1 até *M*5 for diferente de *n+1*, o valor parcial do índice de *Spearman* $(R_{i,MC} - R_{i,MB})^2$ é calculado para o par *i*, do contrário, o valor da correlação é desconsiderado.

Entidades Relacionadas	MB	Ordem	M1	M2	M3	M4	M5	R1	R2	R3	R4	R5
xml	3.240.000	1	6	11	11	11	11	25	100	100	100	100
rdf	3.140.000	2	2	11	11	11	11	0	81	81	81	81
ontology	1.530.000	3	1	11	11	11	11	4	64	64	64	64
networks	1.470.000	4	10	3	11	11	11	36	1	49	49	49
web services	1.460.000	5	4	11	11	11	11	1	36	36	36	36
owl	732.000	6	8	11	11	11	11	4	25	25	25	25
knowledge management	713.000	7	11	6	11	11	6	16	1	16	16	1
agent	623.000	8	5	11	11	11	11	9	9	9	9	9
interoperability	547.000	9	7	1	11	11	11	4	64	4	4	4
information systems	474.000	10	11	8	11	11	11	1	4	1	1	1
environments	471.000	11	11	4	2	4	11	*	49	81	49	*
reasoning	439.000	12	9	11	11	11	11	9	*	*	*	*
patterns	401.000	13	11	9	3	10	1	*	16	100	9	144
daml	302.000	14	3	11	11	11	11	121	*	*	*	*
user interface	262.000	15	11	11	11	6	2	*	*	*	81	169
simulation	253.000	16	11	10	5	11	8	*	36	121	*	64
knowledge representation	237.000	17	11	7	11	11	10	*	100	*	*	49
hypertext	231.000	18	11	11	10	5	3	*	*	64	169	225
intelligent systems	164.000	19	11	11	8	7	11	*	*	121	144	*
trees	157.000	20	11	11	11	9	7	*	*	*	121	169
electronic commerce	140.000	21	11	2	4	11	9	*	361	289	*	144
hypermedia	123.000	22	11	11	11	8	11	*	*	*	196	*
problem solving	118.000	23	11	11	6	11	5	*	*	289	*	324
relational database	83.500	24	11	5	1	2	4	*	361	529	484	400
database management	75.600	25	11	11	11	3	11	*	*	*	484	*
ontology engineering	68.000	26	11	11	9	1	11	*	*	289	625	*
system architecture	44.300	27	11	11	7	11	11	*	*	400	*	*
								230	1308	2668	2747	2058
							Spearman	0,930	0,601	0,186	0,161	0,372

Tabela 31 - Exemplo de cálculo do índice de *Spearman* para a entidade "*Semantic Web*" e os seus pares correlacionados

Nota: MB=mecanismo de busca, ordem=ordem atribuída em função de MB, *M*1 até *M*5 = ordem atribuída pelos modelos de correlação (LRD, *Phi-squared,* MI, VMI e *Z score*) e *R*1 até *R*5 = cálculo parcial do índice de *Spearman* $(R_{i,MC} - R_{i,MB})^2$.

A Tabela 32 apresenta os valores sumarizados para as classes Organização, Pessoa, Área de pesquisa e a média das três classes para o índice de correlação de *Spearman.* O algoritmo LRD atinge os melhores resultados diante dos demais métodos, seguido pelos métodos *Z score* e *Phi-squared*, enquanto que VMI e MI apresentam os piores resultados.

Spearman [-1,1]		Organização	Pessoa	Área de Pesquisa	Média
Sem janela	**LRD**	0,4231	0,4496	0,3979	0,4236
	Phi-squared	0,1487	-0,0291	0,1306	0,0834
	Z score	0,1487	-0,0291	0,1306	0,0834
	MI	0,0797	-0,1525	0,0515	-0,0071
	VMI	0,0797	-0,1525	0,0515	-0,0071
Janela (20)	**LRD**	0,3167	0,1544	0,2793	0,2501
	Z score	0,0029	-0,1393	0,0543	-0,0274
	Phi-squared	-0,0568	-0,1191	-0,0290	-0,0683
	VMI	-0,0472	-0,2083	0,0352	-0,0734
	MI	-0,1310	-0,2283	-0,0606	-0,1400
Janela (50)	**LRD**	0,3286	0,2572	0,2739	0,2866
	Z score	0,1126	-0,0764	0,0231	0,0198
	Phi-squared	0,0598	-0,0562	-0,0180	-0,0048
	VMI	0,0768	-0,1367	0,0173	-0,0142
	MI	0,0157	-0,1715	-0,0542	-0,0700
Janela (100)	**LRD**	0,3423	0,2788	0,3515	0,3242
	Phi-squared	0,1073	-0,0519	0,091	0,0488
	Z score	0,1161	-0,0884	0,0759	0,0345
	VMI	0,099	-0,1344	0,0808	0,0151
	MI	0,0624	-0,1551	0,0339	-0,0196
Janela (200)	**LRD**	0,3847	0,3452	0,398	0,3759
	Phi-squared	0,138	-0,0231	0,162	0,0923
	Z score	0,1075	-0,0844	0,1417	0,0549
	VMI	0,0824	-0,1317	0,132	0,0276
	MI	0,0803	-0,1298	0,0827	0,0111

Tabela 32 - Valores do índice de *Spearman* entre -1 e 1 para as classes Organização, Pessoa e Área de pesquisa, e a média das três classes em diferentes janelas

Analisando-se o Figura 18, verifica-se que isso somente se altera quando considerada a média de todas as janelas para a classe Pessoa. Nesta classe, o método *Phi-squared* se sobressai em relação ao método *Z score*.

À medida que o valor da janela é incrementado, o índice de correlação também o é. Em geral, isso ocorre porque em ambos os modelos a freqüência conjunta é incrementada, promovendo assim uma aproximação melhor diante da ordem estabelecida através do mecanismo de busca. O problema dos modelos MI e VMI se deve à deficiência de lidarem com eventos de baixa freqüência, ou seja, bigramas compostos de palavras com baixa freqüência podem atingir índices maiores do que bigramas compostos de palavras com alta freqüência (MANNING; SCHÜTZE, 1999).

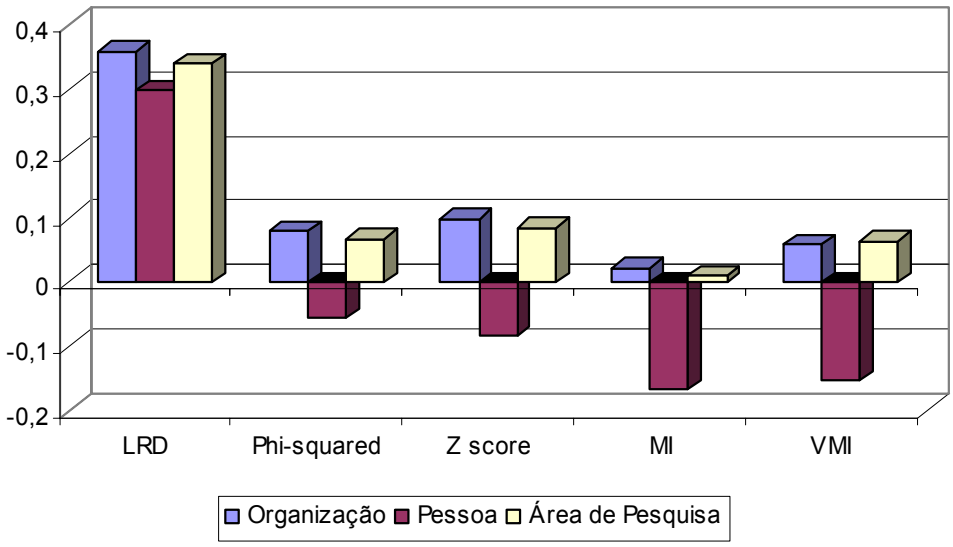


Figura 18 - Correlação média considerando todas as janelas

a) SUMARIZAÇÃO DOS RESULTADOS

A avaliação realizada baseou-se em um conjunto de 300 artigos na área de Web Semântica e possui como objetivo medir o grau de correção entre pares de entidades relacionadas. De modo a evitar análises subjetivas (abordagem orientada a tarefas), as relações entre entidades sugeridas pelos métodos LRD, *Phi-squared*, MI, VMI e *Z score* foram confrontadas com um mecanismo de busca. Como resultado final o algoritmo LRD atinge os melhores resultados diante dos demais métodos, seguido pelos métodos *Z score* e *Phi-squared*, enquanto que VMI e MI apresentam os piores resultados. Novamente, o parâmetro de janela possui pouca influência, e os melhores resultados são obtidos sem qualquer restrição.

4.2 APLICABILIDADE DO MODELO

As fases do modelo proposto discutidas anteriormente, principalmente a extração e a correlação de entidades e expansão do modelo de espaço vetorial, promovem suporte a aplicações de recuperação de informação e à análise de agrupamentos verificando como

entidades e/ou unidade da análise se interconectam, bem como promovem suporte a aplicações de Engenharia e Gestão do Conhecimento.

No âmbito da recuperação de informação, modelos de espaço vetorial são tradicionalmente utilizados na indexação de documentos e na recuperação desses documentos através de consultas baseadas em termos. A expansão vetorial proposta neste trabalho modifica o espaço vetorial com novos termos, ou seja, termos que não estavam no documento original, mas que são relacionados aos existentes. Considerando-se que consultas baseadas em termos são freqüentemente uma aproximação para a informação que se está buscando, os vetores expandidos podem conduzir ao aumento da precisão durante a recuperação de documentos. Nesse sentido, documentos que normalmente não seriam recuperados por um determinado termo ou conjunto de termos, pelo fato de não pertencerem ao espaço vetorial original, passam a ser retornados. Por exemplo, seja o termo $T1$=”Inteligência Artificial” pertencente ao documento $D1$ e o termo $T2$=”Fuzzy” adicionado ao documento $D1$ pelo processo de expansão, uma pesquisa pelo termo $T2$ irá recuperar o documento $D1$, que pode ser útil para a consulta executada pelo usuário.

No contexto de aplicações de agrupamentos, o modelo proposto pode conduzir a análises mais avançadas de como elementos textuais, por exemplo, entidades e conceitos, e mesmo documentos, estão interconectados. A análise desses relacionamentos pode conduzir à identificação de conhecimentos latentes e auxiliar na construção de aplicações de gestão do conhecimento, tornando possível, por exemplo, a apresentação de mapas de conhecimento de uma determinada área de interesse e mesmo promovendo suporte ao estabelecimento de relações diretas e indiretas.

Essas entidades e suas relações representam a base para aplicações de Engenharia e Gestão do Conhecimento, tais como manutenção de ontologias, comunidades de prática, redes sociais, localização de especialistas e gestão por competências. Essas aplicações representam

importantes ferramentas no auxílio ao entendimento das relações que se estabelecem entre entidades, seja de maneira induzida ou espontânea, tais como quem trabalha com quem, em quais projetos, com quais organizações.

No contexto da Engenharia do Conhecimento, ontologias e conseqüentemente sistemas baseados em ontologias e sistemas voltados à manutenção de ontologias desempenham um importante papel (CORCHO et al., 2003; POLI, 2002; LAMMARI; METAIS, 2004). Nesse sentido, entidades e seus relacionamentos servem de insumo, ou seja, provêem instâncias a ontologias. De acordo com Broekstra et al. (2002), ontologia possui papel-chave nos processos de troca de informação. Adicionalmente, ontologia é apresentada como uma possível solução para problemas de compartilhamento de conhecimento (SILVA et al., 2002) ou, como sugerem Uschold e Gruninger (1996), uma ontologia possibilita ainda o mapeamento/integração de modelos de diferentes domínios em uma estrutura coerente.

Muitas definições sobre ontologia têm sido propostas. Russel e Norvig (1995) definem ontologia como uma lista informal de conceitos em um determinado domínio de aplicação. Para Broekstra et al. (2002), ontologia é uma especificação formal e explícita de uma conceitualização compartilhada. Por "conceitualização" entende-se um modelo abstrato de determinado fenômeno que ocorre no mundo real, em que fatos relevantes desse fenômeno são identificados e devem ser explicitamente definidos (BROEKSTRA et al., 2002; GRUBER, 1993; GUARINO; GIARETTA, 1995). O conceito de "formal" refere-se ao fato de que uma ontologia deve ser entendível por agentes de hardware ou software. Por último, o conceito de "compartilhado" reflete a idéia de que uma ontologia captura conhecimento consensual, ou seja, não é restrita a um indivíduo, mas aceita dentro de um grupo.

Finalmente, no contexto da Gestão do Conhecimento, comunidades de prática, localização de especialistas e gestão por competência constituem-se em importantes ferramentas voltadas ao entendimento das relações entre as organizações e os seus

colaboradores e a tomada de decisão. Questões típicas de interesse dos gestores de conhecimento são, por exemplo: (a) quais conhecimentos seus colaboradores possuem; (b) com quais clientes eles têm contato; e (c) quais colaboradores possuem determinados perfis/competências para um projeto em particular. Muitas dessas respostas residem em bases textuais compostas de relatórios técnicos, projetos, e-mails, mensagens instantâneas, etc. e em bases estruturadas ou ontologias organizacionais, que mantêm registradas as interações entre os membros da comunidade nas suas atividades diárias (WENGER, 1998).

De acordo com Alani et al. (2003a), as comunidades de prática (CP) são constituídas por grupos de pessoas com interesses em comum, seja um trabalho ou uma atividade em particular, procedimentos ou domínio de aplicação. Lesser e Stock (2001) ainda definem comunidades de prática como grupos engajados no compartilhamento de experiências e aprendizado através dessas experiências.

Outro exemplo de aplicação são os sistemas de Gestão por Competência. Cada vez mais as organizações estão atentas aos conhecimentos e às habilidades de seus colaboradores por considerarem esses recursos valiosos, isto é, conhecer quem conhece o quê tem se tornado uma atividade crítica, por exemplo, na gerência e na definição de equipes de projetos. A Gestão por Competência objetiva assim o desenvolvimento e a disseminação de competências-chave entre os colaboradores da organização. Como declarado por Hamel e Prahalad (1994), “competências-chave transcendem produtos ou serviços em particular e de fato, podem transcender qualquer departamento dentro de uma organização”. De acordo com Dawson (1991), competências-chave constituem-se na combinação de habilidades já aprendidas ou em desenvolvimento, podendo ainda fomentar ou determinar estratégias de negócio (HAFEEZ et al., 2002). Assim, métodos utilizados na descoberta ou na avaliação de competências-chave têm se mostrado relevantes para as organizações.

As aplicações discutidas acima são exemplos de onde o trabalho se insere, pois possuem como fonte elementos textuais e seus relacionamentos. Tal informação pode ser adquirida através de sistemas específicos com o objetivo de coletar e sumarizar informações pessoais. Um exemplo disso é a coleção de currículos de pesquisadores brasileiros mantidos através da Plataforma Lattes (CNPq, 2005). Podem, entidade e seus relacionamentos, igualmente advir de documentos em geral, tais como páginas *Web* e relatórios, que refletem atividades diárias dentro de uma organização, ou mesmo comunicações persistentes, compostas de mensagens eletrônicas ou registros de comunicações entre os colaboradores.

Com o intuito de explorar e auxiliar no entendimento dos relacionamentos estabelecidos entre entidades em determinado contexto, como, por exemplo, em uma organização, uma coleção de documentos ou um domínio *Web*, apresenta-se uma ferramenta com esse propósito.

A Figura 19 mostra a ferramenta por meio da qual, a partir de uma entidade utilizada na consulta, são recuperadas as entidades mais relacionadas, considerando-se algumas classes, entre elas, Organização, Pessoa, Projeto e Área de pesquisa. As entidades recuperadas são classificadas/ordenadas levando-se em conta o grau de relacionamento com a entidade de origem. Essa abordagem possibilita rapidamente a inspeção das relações mais importantes de cada entidade em cada classe e atende de maneira prática à investigação de competências associadas a determinada pessoa ou mesmo à localização de pessoas em um determinado assunto (área de pesquisa).

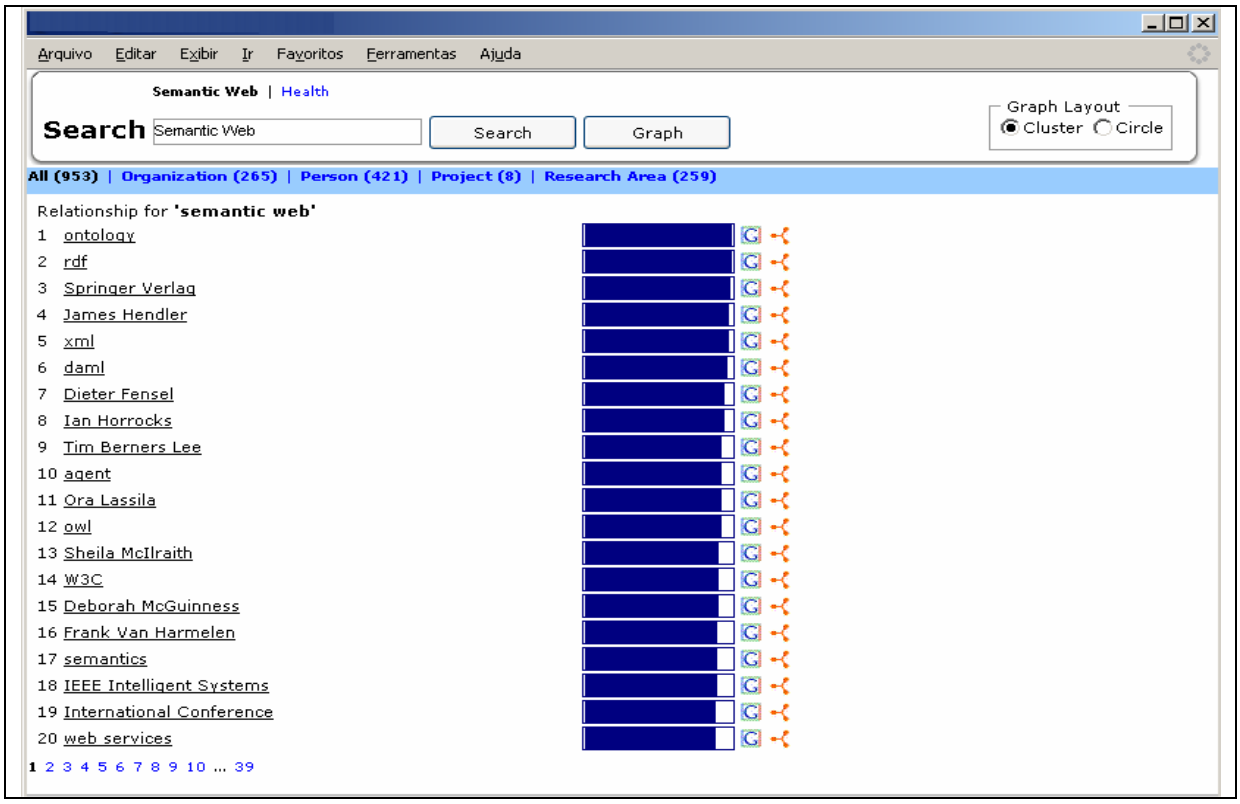


Figura 19 - Aplicação responsável pela apresentação de entidades e seus relacionamentos

Essa abordagem pode ser incrementada apresentando-se diferentes níveis de relacionamentos. Isso ocorre através da utilização de agrupamentos com o objetivo de explicitar relações latentes entre entidades. Nesse processo utilizam-se os vetores extraídos a partir da matriz de correlação, similar à matriz apresentada na Tabela 22. Cada vetor representa uma determinada entidade. Ao final do processo, descrito na Seção 3.5, a entidade que melhor se adere ao agrupamento irá indicar o centro dele e as demais entidades do agrupamento serão relacionadas à entidade principal, formando um mapa de relacionamentos.

Isso é demonstrado através da utilização de representação gráfica (Figura 20), possibilitando assim um melhor entendimento dos relacionamentos entre entidades. De maneira similar à versão textual, diferentes classes são utilizadas e identificadas com diferentes cores. Na figura a seguir a forma retangular indica a entidade que melhor representa o centro de um determinado agrupamento, enquanto que a forma circular indica as entidades relacionadas. Como mencionado anteriormente, entidades são identificadas através do processo chamado NER, que se utiliza de uma base de conhecimento para nomear/atribuir um

determinado elemento textual para uma classe de entidade específica. Essa base de conhecimento possui, para cada classe do sistema, por exemplo, pessoa, organização, projeto, uma tabela de nomes associada, permitindo assim a classificação desses elementos textuais.

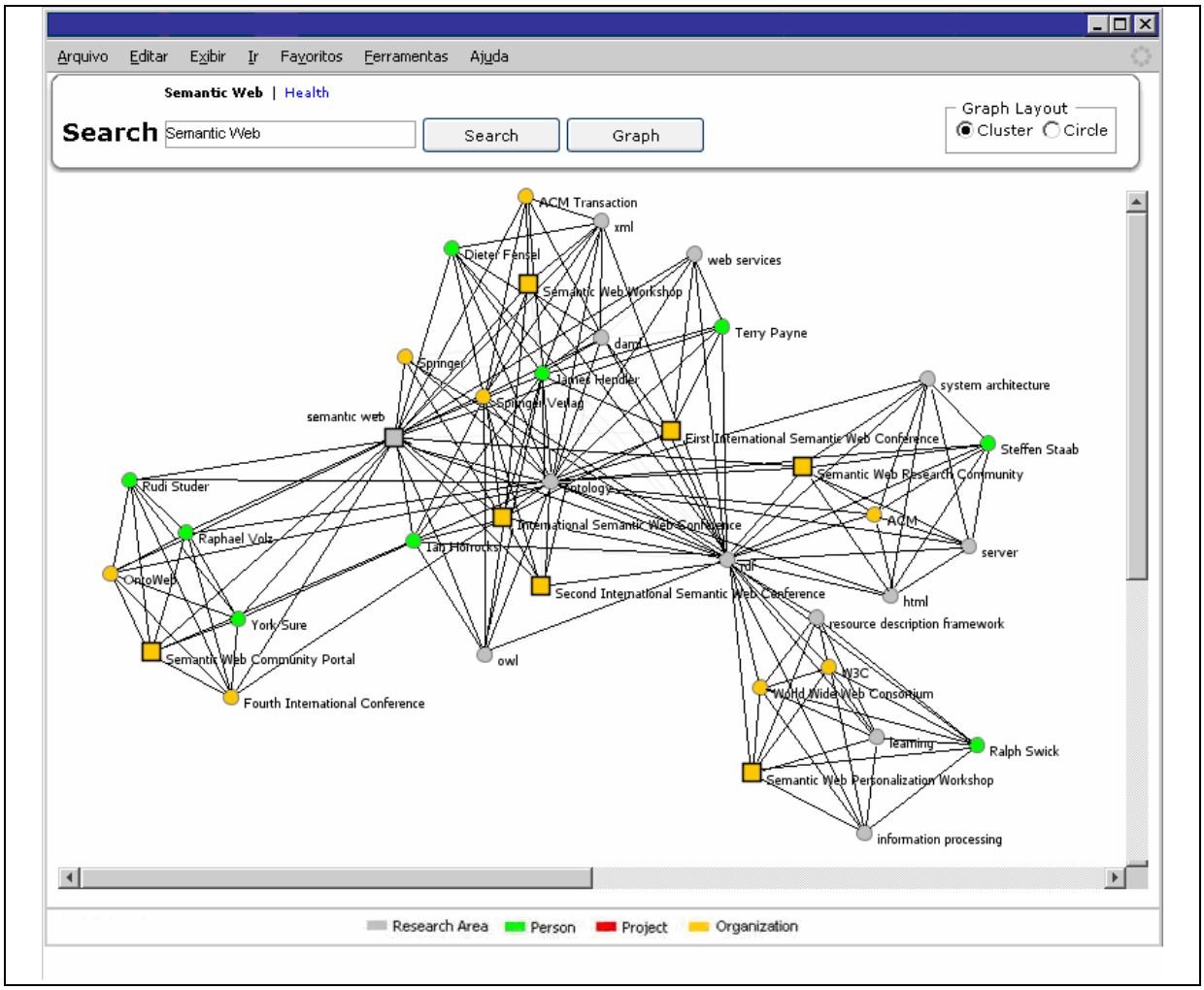


Figura 20 - Aplicação responsável pela apresentação gráfica de agrupamentos de entidades e seus relacionamentos

4.3 CONSIDERAÇÕES FINAIS

Este capítulo discutiu a validação do modelo proposto (LRD), aplicando-se a ele as tarefas de recuperação de informação e agrupamentos, bem como a avaliação da capacidade de identificação de relacionamentos relevantes entre elementos textuais. No contexto da recuperação de informação, foram empregadas as medidas de precisão e lembrança, tradicionalmente utilizadas na mensuração do desempenho de sistemas dessa natureza. Para agrupamentos utilizou-se o erro quadrático com o objetivo de determinar o desempenho do

modelo nessas tarefas. No último conjunto de avaliações empregou-se um coeficiente de correlação, visando estabelecer a habilidade do modelo em informar relevantes listas de entidades relacionadas a determinado elemento textual (entidade ou conceito). Por fim, discutiu-se a aplicabilidade do modelo nos cenários de recuperação de informação, agrupamento e aplicações de Engenharia e Gestão do conhecimento. De modo geral entidades e suas relações formam vetores de conhecimento e oferecem suporte a aplicações de análises de comunidades de prática, redes sociais, gestão por competências, localização de especialistas, manutenção de ontologias, entre outras.

5 CONCLUSÕES E TRABALHOS FUTUROS

A descoberta de que o universo está em expansão foi uma das grandes revoluções intelectuais do século XX. Depois dela torna-se fácil perguntar por que ninguém pensou nisso antes.

Stephen W. Hawking

5.1 CONCLUSÕES

O conhecimento tem se tornado fator-chave para as organizações. Com o crescimento da *Web* e das *Intranets* organizacionais, milhares de documentos estão atualmente disponíveis. Embora muitas organizações mantenham estruturas padronizadas de dados, tais como modelos relacionais ou ontologias representando as atividades operacionais da organização, esses modelos tendem a ser mais estáticos e com maior custo de manutenção. Basicamente refletem o que deveria acontecer na organização, e não necessariamente a realidade do dia-a-dia. Essa realidade é mais bem caracterizada em documentos, tais como relatórios técnicos, artigos, e-mails, registros de conversas *on-line*, etc. Através da análise do conteúdo textual é possível a identificação/extração de entidades/conceitos e o estabelecimento da força da relação (correlação) entre esses elementos. Como resultados desse processo surgem modelos que auxiliam no entendimento das inter-relações entre elementos textuais (entidades/conceitos).

O entendimento dessas relações possibilita, por exemplo, o compartilhamento e/ou a disseminação do conhecimento, a verificação de interesses de pesquisa comuns e espontâneos, a análise de redes de colaboração intra e interorganização, a definição de equipes de trabalho e a localização de especialistas. Nesse sentido, o desafio reside em como extrair e correlacionar relevantes padrões textuais de modo que isso promova a descoberta de

conhecimento latente e mesmo possibilite uma melhor gestão do conhecimento em benefício da organização.

O modelo proposto neste trabalho chamado LRD (*Latent Relation Discovery*) integra a extração e a correlação de elementos textuais (entidades e conceitos) e, através da recuperação de informação e da descoberta de conhecimento, visa promover suporte a aplicações de Engenharia e Gestão do Conhecimento. Como resultado do processo, matrizes de correlação são produzidas. No âmbito da recuperação de informação e de agrupamentos, tais matrizes são insumos para expansão do espaço vetorial. Já no âmbito de aplicações de gestão do conhecimento, possibilitam uma rápida identificação dos relacionamentos diretos entre entidades/conceitos. Permitem ainda, através da aplicação de agrupamentos, a identificação de relacionamentos indiretos. Além disso, cada entidade/conceito e seus relacionamentos formam vetores de conhecimento que indicam o perfil/contexto de determinado elemento.

No cerne do modelo encontram-se a correlação de elementos textuais e a expansão de unidades de análise. Com o objetivo de validar o modelo proposto, foram realizados testes quantitativos, baseados em conjunto de dados padronizados e orientados a tarefas.

Nos dois primeiros testes as relações entre os elementos textuais foram utilizadas para incrementar o contexto de representação vetorial tradicional de documentos de modo a fornecer adicional significado e melhorar tanto a recuperação de informação quanto o agrupamento de documentos.

Considerando-se o primeiro experimento baseado no conjunto de dados CISI, LRD demonstrou um incremento na medida *F* quando aplicado ao modelo de espaço vetorial tradicional, obtendo melhores resultados se comparado a outros cinco métodos (LSI, *Phi-squared*, MI, VMI e *Z score*). Isso pode ser observado tanto na avaliação utilizando a média quanto na avaliação com o limiar de 0.5.

Para o segundo experimento baseado no conjunto de dados REUTERS, LRD demonstrou um melhor desempenho ao incrementar o contexto de documentos utilizados na tarefa de agrupamento. Os resultados obtidos através do método LRD foram superiores a outros quatros métodos, *Phi-squared*, MI, VMI e *Z score*, tanto na avaliação utilizando a média de todos os limiares e janelas, quanto na avaliação levando-se em conta o limiar de 0.5.

Além desses experimentos, realizaram-se mais dois orientados à tarefa e outro quantitativo visando medir a habilidade do método LRD em fornecer relevantes listas de entidades relacionadas a determinada consulta. LRD demonstrou resultados melhores quando comparado aos métodos *Phi-squared*, MI, VMI e *Z score.* A relevância desses resultados reside no fato de que listas de entidades/conceitos ordenadas de maneira adequada promovem melhor suporte às aplicações de Gestão do Conhecimento.

Como declarado anteriormente, os resultados dos experimentos estão limitados aos conjuntos de dados utilizados nas avaliações. Outra limitação refere-se à possibilidade de múltiplos sentidos para uma determinada entidade. Apesar de o modelo ser aplicável a qualquer coleção de documentos, a falta do tratamento de múltiplos sentidos de maneira automática ou não pode conduzir a erros no estabelecimento na força do relacionamento entre entidades.

Ressalta-se ainda que a utilização de agrupamentos vai ao encontro da fase de visualização proposta no modelo, de modo a produzir meios de se analisarem entidades e seus relacionamentos através de mapas de conhecimento. Entretanto, outros tipos de conhecimento poderiam ser utilizados, tais como regras de associação ou classificação.

Finalmente, menciona-se que LRD e LSI possuem comportamento similar, uma vez que ambos os modelos podem lidar com comparações termo–termo, documento–documento ou termo–documento. Todavia, LSI requer tempo elevado de processamento quando aplicada a grandes coleções de documentos (IKEHARA et al., 2001). Por outro lado, os experimentos

com os conjuntos de dados CISI e REUTERS mostram que o tempo de execução do método, ou seja, o tempo necessário para correlacionar elementos textuais, tende a aumentar linearmente em função do número de entidades e documentos examinados. O método pode ainda reavaliar relações já estabelecidas assim como estabelecer novas relações incrementalmente à medida que novos documentos são incorporados.

5.2 TRABALHOS FUTUROS

Como trabalhos futuros vislumbram-se tanto o melhoramento do modelo atual quanto a sua utilização em diversas aplicações.

Primeiramente, objetivam-se o estudo e o desenvolvimento de novas métricas voltadas ao estabelecimento das relações entre elementos textuais. Além da precisão, o desempenho é um fator crítico. Apesar dos resultados obtidos com o modelo atual, uma versão distribuída poderá incrementar em muito a escalabilidade e o desempenho quando aplicada a grandes coleções de documentos. Nesse sentido, avaliações sobre as coleções de documentos TREC são pretendidas. Um modelo probabilístico ou neural capaz de predizer a força de relacionamento entre elementos textuais, considerando somente os *n* documentos mais importantes em que um par de entidades/conceitos ocorre, também denota perspectivas de desenvolvimento e de pesquisa.

No modelo atual a extração de entidades/conceitos ocorre através do processo de NER ou da utilização de dicionários controlados/*thesaurus*/ontologias. Um modelo incremental baseado no registro das consultas efetuadas em coleções de documentos permitiria capturar o interesse do usuário. Por “interesse” entende-se todo e qualquer argumento de busca utilizado pelo usuário na localização de documentos. Sendo assim, cada registro de consulta (interesse do usuário) pode conter um ou mais termos que, por meio de análises adequadas, podem ser validados como conceitos e/ou entidades. O registro desses elementos, assim como a

identificação dos documentos nos quais eles ocorrem, possibilita a correlação e a conseqüente criação de matrizes de correlação.

Nesse sentido, o modelo incremental é mais dinâmico e possui como principal vantagem a não dependência de bases de conhecimento no suporte ao processo de extração de informação quando somente a identificação de conceitos é requerida. Cabe ressaltar que a fase de identificação de entidades necessita de uma base de conhecimento capaz de mapear as várias classes consideradas no processo. De modo geral, esses três modelos podem operar de maneira complementar, objetivando auxiliar na extração automática de elementos textuais.

Ressalta-se ainda que elementos textuais extraídos a partir de textos representam instâncias genéricas de uma ontologia, assim como a força do relacionamento entre esses elementos estabelece o peso de uma ou várias relações. Nesse sentido, são desejáveis aplicações capazes de organizar e classificar elementos textuais, e estabelecer tipos de relacionamento entre os elementos. Assim, um processo semi-automático de manutenção de ontologia pode incrementar tanto a recuperação de informação quanto aplicações de Engenharia e Gestão do Conhecimento.

Como apresentado na Seção 4.2, os mapas de conhecimento possibilitam análises mais apuradas de entidades/conceitos e de seus relacionamentos, tendo impacto em aplicações como análise de comunidades de prática e redes sociais. Entretanto, com base no modelo atual, somente é possível gerar a força do relacionamento, e não o(s) tipo(s) do(s) relacionamento(s). Para tal, ontologias fornecem meios para mapear os diversos tipos de relações entre instâncias, tornando-se particularmente útil na análise e nos entendimentos das relações estabelecidas entre entidades.

Outra aplicabilidade do modelo seria em ambientes de aprendizado. Matrizes de correlação podem fornecer subsídios para incrementar o perfil dos participantes do ambiente. A redefinição ou o melhoramento do contexto original de determinado perfil pode auxiliar na

indicação de outros participantes com perfis similares, possibilitando a formação de redes de colaboração entre os participantes do ambiente. Adicionalmente, a análise do perfil e do registro de conteúdo acessado pode sugerir material adicional (artigos, livros, etc.) que determinado participante deveria estudar.

Finalmente, mas não de maneira exaustiva, vislumbra-se o emprego da extração e da correlação de conceitos com posterior projeção através de mapas de conhecimento em ambientes colaborativos de disseminação de informação, como, por exemplo, os *wikis*. Tal aplicação poderia possibilitar um melhor entendimento do modo como os conceitos se relacionam, facilitando o acesso ao conteúdo disponibilizado pelo ambiente. Por outro lado, a utilização do ambiente sugere indícios sobre o interesse de determinado usuário. Nesse sentido a aplicação no modelo constitui-se como uma ferramenta em potencial para a disseminação de conhecimento, uma vez que permite mapear comunidades de prática constituídas de maneira espontânea.

REFERÊNCIAS BIBLIOGRÁFICAS

AGICHTEIN, E.; GRAVANO, L. Snowball: Extracting Relations from Large Plain-Text Collections. In: ACM INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES, 5., 2000. **Proceedings…** 2000, p. 85–94.

ALANI, H.; DASMAHAPATRA, S.; O'HARA, K.; SHADBOLT, N. Identifying communities of practice through ontology network analysis. **IEEE Intelligent Systems**, v. 18, n. 2, p. 18-25, 2003a.

ALANI, H.; KIM, S.; MILLARD, D.E.; WEAL, M.J.; HALL, W.; LEWIS, P.H.; SHADBOLT, N.R. Automatic ontology-based knowledge extraction from web documents. **IEEE Intelligent Systems**, v. 18, n. 1, p. 14-21, 2003b.

ALAVI, M; COOK, J; COOK, L; LEIDNER, D.E. Review: Knowledge Management and knowledge management systems: Conceptual foundations and research issues. **MIS Quarterly**, v. 25, n. 1, p. 107-136, 2001.

ANDREASEN, T.; BULSKOV, H.; KNAPPE, R. On ontology-based quering. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, ONTOLOGIES AND DISTRIBUTED SYSTEMS, 18., 2003. **Proceedings…** 2003. p. 53-59.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: Addison-Wesley, 1999.

BERRY, M.J.A.; LINOFF, G. **Data mining techniques - for marketing, sales, and customer support**. New York: John Wiley & Sons, 1997.

BI, Y.; ANDERSON, T.; MCCLEAN, S. A rough set model with ontologies for discovering maximal association rules in document collections. **Knowledge-Based Systems**, v. 16, n. 5-6, p. 243-251, 2003.

BIGUS, Joseph P. **Data mining with neural networks: Solving business problems from application development to decision support**. Computing McGraw-Hill, New York, NY, 1996.

BILLHARDT, H.; BORRAJO, D.; MAOJO, V. A context vector model for information retrieval. **Journal of the American Society for Information Science and Technology**, v. 53, n. 3, p. 236-249, 2002.

BONTCHEVA, K.; TABLAN, V.; MAYNARD, D.; CUNNINGHAM, H. Evolving GATE to meet challenges in language engineering. **Natural Language Engineering**, v. 10, n. 3-4, p. 349-373, 2004.

BOOCH, G.; RUMBAUGH, J; JACOBSON, I. The Unified Modelling Language User Guide, Massachusetts: Addison Wesley, 1999.

BRIN, S. Extracting Patterns and Relations from the World Wide Web. In: WEBDB, 1998. **Proceedings…,** 1998, p. 172-183.

BROEKSTRA, J.; KLEIN, M.; DECKER, S.; FENSEL, D.; HARMELEN, F. van; HORROCKS, I. Enabling knowledge representation on the web by extending RDF Schema. **Computer Networks**, v. 39, n. 5, p. 609-634, Aug. 2002.

CALIFF, M. E.; MOONEY, R. J. Relational learning of pattern-match rules for information extraction. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-99), 16., 1999, Orlando, Florida. **Proceedings…** Orlando, Florida, 1999, p. 328–334.

CARDIE, Claire. Empirical methods in information extraction. **AI Magazine**, v. 18, n. 4, p. 65-79, 1997

CARPENTER, G. A.; GROSSBERG, S. A massively parallel architecture for a self-organizing neural pattern recognition machine. **Computer Vision, Graphics, and Image Processing**, v. 37, p. 54-115, 1987.

CASTILLO, G.; SIERRA, G.; MCNAUGHT, J. An improved algorithm for semantic clustering. In: INTERNATIONAL SYMPOSIUM ON INFORMATION AND COMMUNICATION TECHNOLOGIES, ACM International Conference Proceeding Series, 1., 2003, Dublin, Ireland. **Proceedings...** Dublin, Ireland, 2003. p. 304-309.

CHEN, M-S.; HAN, J.; YU, P.S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n.6, p. 866-883, 1996.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational Linguistics**, v. 16, n. 1, p. 22-29, 1990.

CHURCH, Kenneth W.; GALE, William A. Concordances for Parallel Text. In: ANNUAL CONFERENCE OF THE UW CENTRE FOR THE NEW OED AND TEXT RESEARCH, 7., 1991, Oxford , England. **Proceedings…** Oxford, England, 1991, p. 40-62.

CHURCH, Kenneth W.; MERCER, Robert L. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. **Computational Linguistics**, v. 19, pp. 1-24, 1993.

CIRAVEGNA, F. Adaptive Information Extraction from Text by Rule Induction and Generalisation. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI), 17., 2001. Seattle, USA. **Proceedings...** Seattle, USA, 2001.

CIRAVEGNA, F; WILKS, Y. Designing adaptive information extraction for the semantic web in amilcare. In: HANDSCHUH, S.; STAAB, S. (Ed.). ANNOTATION FOR THE SEMANTIC WEB, FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS, IOS Press, Amsterdam, 2003.

CNPq. **Historic of the Lattes Platform**. 2005. Disponível em: <http://lattes.cnpq.br/pl/curriculo/historico.jsp>. Accessado em: 10 de Maio de 2005.

CONRAD, Jack G.; UTT, Mary Hunter. A System for Discovering Relationships by Feature Extraction from Text Databases. SIGIR, pp. 260-270, 1994.

CORCHO, O.; FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A. Methodologies, tools and languages for building ontologies. Where is their meeting point?. **Data & Knowledge Engineering**, n. 46, p. 41-64, 2003.

CROFT, W. B.; TURTLE, H. R.; LEWIS D. D. The use of phrases and structured queries in information retrieval. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR '91), 14., 1991, New York, USA. **Proceedings…** New York: ACM Press, 1991, p. 32-45.

CUNNINGHAM, H. GATE: a General Architecture for Text Engineering. **Computers and the Humanities**, v. 36, p. 223-254, 2002.

DARPA - Defense Advanced Research Projects Agency. In: MESSAGE UNDERSTANDING CONFERENCE, 6., 1995. **Proceedings…** Morgan Kaufmann, 1995.

DAVENPORT, T.H.; PRUSAK, L. **Information ecology**: Mastering the information and knowledge environment, Oxford University Press, 1997.

DAVENPORT, T.H.; PRUSAK, L. **Working knowledge**: How organizations manage what they know, Harvard Business School Press, Boston, 2000.

DAWSON, K. Core competency management in R&D organizations. In: KOCAOGLU, D.; NIWA, K. (Ed.). **Technology Management**: The New International Language, New York: Institute of Electrical and Electronics Engineers, 1991. p. 145-148.

DEERWESTER, S. C.; DUMAIS, S. T.; LANDAUER, T. K.; FURNAS, G. W.; HARSHMAN, R.A. Indexing by latent semantic analysis. **Journal of the American Society of Information Science**, v. 41, n. 6, p.391-407, 1990.

DING, C.; HE, X. K-means clustering via principal component analysis. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML 2004), 21. 2004. **Proceedings…** 2004. p. 225-232.

DING, C.H.Q. A probabilistic model for dimensionality reduction in information retrieval and filtering. In: SIAM COMPUTATIONAL INFORMATION RETRIEVAL WORKSHOP, 1., 2000, Raleigh, NC. **Proceedings…** Raleigh, NC, 2000.

DÖRRE, J.; GERSTI, P.; SEIFFERT, R. Text mining: Finding nuggets in mountains of textual data. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 5. 1999. **Proceedings…** 1999, p. 398-401.

DOZIER, C.; JACKSON, P., GUO, X.; CHAUDHARY, M.; ARUMAINAYAGAM, Y. Creation of an Expert Witness Database through Text Mining. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW, 9., 2003, **Proceedings…** 2003, p. 177-184.

DRUCKER, P. **Managing for the future**: The 1990's and beyond. New York: Truman Tally Books/Dulton, 1992.

DUDA, R.; HART, P. Pattern classification and scene analysis, New York: Wiley, 1973.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. **Information Processing and Management: an International Journal**, v. 38, n. 6, p. 823-848, 2002.

ETZIONI, O.; CAFARELLA, M.; DOWNEY, D.; POPESCU, A.; SHAKED, T.; SODERLAND, S.; WELD, S.; YATES, A. Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In: AAAI 2004, **Proceedings…** 2004, p. 391-398.

FAYYAD, U. M. Data mining and knowledge discovery: making sense out of data. **IEEE Intelligent Systems**, v. 11, n. 5, p. 20-25, 1996a.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: An overview. In: ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, Cambridge, Massachusetts, and London, England: AAAI Press/The MIT Press, MIT, 1996b. p. 1-34.

FELDMAN, R.; FRESKO, M.; HIRSH, H.; AUMANN, Y.; LIPHSTAT, O.; SCHLER, Y.; RAJMAN, M. Knowledge management: A text mining approach. In: INTERNATIONAL CONFERENCE ON PRACTICAL ASPECTS OF KNOWLEDGE MANAGEMENT (PAKM-98), 2., 1998, Basel, Switzerland. **Proceedings…** Basel, Switzerland, 1998, p. 9.1–9.10.

FORSYTHE, G. E.; MALCOLM, M. A.; MOLER, C. B. Computer methods for mathematical computations (Chapter 9: Least squares and the singular value decomposition). Englewood Cliffs, NJ: Prentice Hall, 1977.

FREITAG, D.; KUSHMERICK, N. Boosted wrapper induction. In: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-2000), 17., 2000, Austin, Texas. **Proceedings…** Austin, Texas, 2000, p. 577–583.

GONÇALVES, A.; UREN, V.; KERN, V.; PACHECO, R. Mining knowledge from textual databases: An approach using ontology-based context vectors. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS (AIA-2005), Innsbruck, Austria. **Proceedings…** Innsbruck, Austria, 2005, p. 66-71.

GONÇALVES, A; BEPPLER, F; BOVO, A; KERN, V; PACHECO, R. A Text Mining Approach towards Knowledge Management Applications. In: INTERNATIONAL WORKSHOP ON INFORMATION RETRIEVAL ON CURRENT RESEARCH INFORMATION SYSTEMS (CRIS-IR 2006), Copenhagen, Denmark. **Proceedings…** Copenhagen, Denmark, 2006a. p. 7-28.

GONÇALVES, A; ZHU, J; SONG, D; UREN, V; PACHECO, R. LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval. In: INTERNATION CONFERENCE ON WEB-AGE INFORMATION MANAGEMENT (WAIM 2006), 7., 2006, Hong Kong. J.X. Yu, M. Kitsuregawa, and H.V. Leong (Eds.): WAIM 2006, Lecture Notes in Computer Science. Springer Verlag. **Proceedings…** Hong Kong, 2006b. v. 4016, p. 122-133.

GRABMEIER, J.; RUDOLPH, A.Techniques of Cluster Algorithms in Data Mining. **Data Mining and Knowledge Discovery**, v.6, p. 303–360, 2002

GROVER, C.; GEARAILT, D. N.; KARKALETSIS, V.; FARMAKIOTOU, D.; PAZIENZA, M. T.; VINDIGNI, M. Multilingual XML-Based Named Entity Recognition for E-Retail Domains. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2002), 3., 2002, Las Palmas. **Proceedings…** Las Palmas, 2002, p. 1060-1067

GRUBER, T.R. Towards principles for the design of ontologies used for knowledge sharing. **International Journal of Human-Computer Studies**, v. 43, p. 907-927, 1993.

GRUHL, D.; CHAVET, L.; GIBSON, D.; MEYER, J.; PATTANAYAK, P.; TOMKINS, A.; ZIEN, J. How to build a WebFountain: An architecture for very large-scale text analytics. **IBM Systems Journal**, v. 43, n. 1, p. 64-77, 2004.

GUARINO, N.; GIARETTA, P. Ontologies and knowledge bases: towards a terminological clarification. In: MARS, N. (Ed.). **Towards Very Large Knowledge Bases**: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 1995, p. 25-32.

GUTHRIE, L.; PUSTEJOWSKY, J.; WILKS, Y.; SLATOR, B. M. The Role of Lexicons in Natural Language Processing. **CACM**, v. 39, n. 1, p. 63-72, 1996.

HAFEEZ, K.; ZHANG, Y.; MALAK, N. Identifying core competence. **IEEE Potencials**, v. 49, n. 1, p. 2-8, 2002.

HAIR Jr., J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Multivariate data analysis**. 5. ed., New Jersey: Prentice-Hall, Upper Saddle River, 1998.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17 n. 2-3, p. 107-145, 2001.

HAMEL, G.; PRAHALAD, C.K. **Competing for the future**. Harvard Business School, Boston, 1994, 352p.

HARRISON, T.H. **Intranet data warehouse**. São Paulo: Berkeley Brasil, 1998.

HEFLIN, J., HENDLER, J. Searching the Web with Shoe. In: AAAI WORKSHOP ON AI FOR WEB SEARCH, 2000.

HERLOCKER, J.L.; KONSTAN, J.A.; TERVEEN, L.G.; RIEDL, J.T. Evaluating collaborative filtering recommender systems. **ACM Transactions on Information Systems**, v. 22, n. 1, p. 5-53, 2004.

HOLSHEIMER, M.; SIEBES, A. **Data mining**: The search for knowledge in databases. Amsterdam: Computer Science/Department of Algorithmics and Architecture, 1994. Disponível em: <http://www.cwi.nl/cwi/publications/index.html>. Acessado em: 17 de Agosto de 1999.

HOTHO, A.; MAEDCHE, A.; STAAB, S. Text clustering based on good aggregations. In: THE 2001 IEEE INTERNATIONAL CONFERENCE ON DATA MINING, IEEE Computer Society, 2001. **Proceedings…** 2001, p. 607-608.

HOTHO, A.; STUMME, G. Conceptual clustering of text clusters. In: FACHGRUPPENTREFFEN MASCHINELLES LERNEN (FGML), 2002, Hannover. **Proceedings…** 2002. p. 37-45.

IDE, N.; VÉRONIS, J. Word Sense Disambiguation: The state of the arte. **Computational Linguistics**, v. 24, n. 1, p. 1-41, 1998.

IKEHARA, S.; MURAKAMI, J.; KIMOTO, Y.; ARAKI, T. Vector space model based on semantic attributes of words. In: PACIFIC ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (PACLING), 2001, Kitakyushu, Japan. **Proceedings…** Kitakyushu, Japan, 2001.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys** (CSUR), v. 31, n. 3, p. 264-323, 1999.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. New Jersey: Prentice-Hall, 4. ed., 1998.

JONES, W.P.; FURNAS, G.W. Pictures of relevance: A geometric analysis of similarity measures. **Journal of American Society for Information Science**, v. 38, n. 6, p. 420-442, 1987.

JUSTESON, J. S.; KATZ, S. M. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. **Natural Language Engineering**, v. 1, pp. 9-27, 1995.

KOHONEN, T. Self-Organizing maps. Springer Series in Information Sciences, Heidelberg, Germany: Springer-Verlag, 1995.

KOHONEN, T.; KASKI, S.; LAGUS, K.; SALOJÄRVI, J.; HONKELA, J.; PAATERO, V. SAARELA, A. Self organization of a massive document collection. IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery, v. 11, n. 3, p. 574-585, 2000.

KORFHAGE, R.R. **Information storage and retrieval**. New York: Wiley Computer Publishing, 1997.

LAMMARI, N.; MÉTAIS, E. Building and maintaining ontologies: a set of algorithms. **Data & Knowledge Engineering**, v. 48, n. 2, p.155-176, 2004.

LANDAUER, T.K.; DUMAIS, S.T. Solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. **Psychological Review**, v. 104, n. 2, p. 211-240, 1997

LEHNERT, W.; CARDIE, C.; FISHER, D.; MCCARTHY, J.; RILOFF, E.; SODERLAND, S. 1992. University of Massachusetts: Description of the CIRCUS System as Used in MUC-4. In: MESSAGE-UNDERSTANDING CONFERENCE (MUC-4), 4., 1992, San Francisco, California. **Proceedings…** San Francisco, California: Morgan Kaufmann, 1992, p. 282–288.

LESSER, E.L.; STORCK, J. Communities of practice and organizational performance. **IBM Systems Journal**, v. 40, n. 4, p. 831-841, 2001.

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. de. Concept-based knowledge discovery in texts extracted from the Web. **SIGKDD Explorations**, v. 2, n. 1, p. 29–39, July 2000.

MACK, R.; HEHENBERGER, M. Text-based knowledge discovery: search and mining of life-sciences documents. **Drug Discovery Today**, v. 7, n. 11, p. S89-S98, 2002.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: BERKELEY SYMPOSIUM OF MATHEMATICS, STATISTICS AND PROBABILITIES, 5., 1967. **Proceedings…** 1967.

MANNING, C.D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. The MIT Press, Cambridge, Massachusetts, 1999.

MARWICK, A.D. Knowledge management technology. **IBM Systems Journal**, v. 40, n. 4, p. 814-830, 2001.

MEDLER, D.A. A brief history of connectionism. **Neural Computing Survey**, v. 1, p. 61-101, 1998.

MILLER, G.; BECKWITH, R.; FELLBAUM, C.; GROSS, D.; MILLER, K. Introduction to WordNet: an on-line lexical database. **International Journal of Lexicography**, v. 3, n. 4, p.235-244, 1990.

MITRA, M.; CHAUDHURI, B.B. Information retrieval from documents: A survey. **Information Retrieval**, n. 2, p. 141-163, 2000.

MOONEY, Raymond J.; NAHM, Un Yong. Text Mining with Information Extraction. In: INTERNATIONAL MIDP COLLOQUIUM DAELEMANS, 4., September 2003, Bloemfontein, South Africa. W., du PLESSIS, T., SNYMAN, C. and TECK, L. (Eds.). **Proceedings…** Bloemfontein, South Africa: Van Schaik Pub., 2005. p.141-160.

MOTTA, E. The knowledge modelling paradigm in knowledge engineering. Handbook of Software Engineering and Knowledge Engineering, **World Scientific Publishing**, 2000.

NAHM, Un Yong; MOONEY, Raymond J. Using Soft-Matching Mined Rules to Improve Information Extraction. In: AAAI-2004 WORKSHOP ON ADAPTIVE TEXT EXTRACTION AND MINING (ATEM-2004), 2004, San Jose, CA. **Proceedings…** San Jose, CA, 2004. p. 27-32.

NASUKAWA, T.; NAGANO, T. Text analysis and knowledge mining system. **IBM Systems Journal**, v. 40, n. 4, p. 967-984, 2001.

NONAKA, I.; TAKEUCHI, H. **The knowledge-creating company**: How japanese companies create the dynamics of innovation, Oxford, UK: Oxford University Press, 1995.

NOUALI, O.; BLACHE, P. A semantic vector space and features-based approach for automatic information filtering. **Expert Systems with Applications**, v. 26, n. 2, p. 171-179, 2003.

PANDYA, A.S.; MACY, R.B. Pattern recognition with neural networks in C++. **CRC Press**, Boca Raton, Florida, Florida Atlantic University, 1995.

POLI, R. Ontological methodology. **Journal of Human-Computer Studies**, v. 56, p. 639-664, 2002.

PPGEP – Programa de Pós-Graduação em Engenharia de Produção. **Proposta de Concepção do Curso**. 2006. Disponível em: < http://www.ppgep.ufsc.br/1a.htm>. Accessado em: 10 de Maio de 2006.

RAMONI, M.; SEBASTIANI, P.; COHEN, P. Bayesian clustering by dynamics. **Machine Learning**, v. 47, p. 91-121, 2002.

RESNIK, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. **Journal of Artificial Intelligence Research** (An International Electronic and Print Journal), v. 11, p. 95-130, 1999.

RILOFF, E. Automatically Constructing a Dictionary for Information-Extraction Tasks. In NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), 11., 1993, Menlo Park, CA. **Proceedings…** Menlo Park, CA, 1993. p. 811–816.

RUSSEL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. Prentice-Hall: New Jersey, 1995. 932p.

SALTON, G.; BUCKLEY, C. Therm-weighting approachs in automatic text retrieval. **Information Processing & Management**, v. 24, n. 5, p. 512-523, 1988.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, v. 18, n. 11, p. 613–620, 1975.

SAVOLAINEN, T.; BEECKMANN, D.; GROUMPOS, P; AND JAGDEV, H. Positioning of modelling approaches, methods and tools. **Computers in Industry**, v. 25, p. 255-262, 1995.

SCHREIBER, G. ; WIELINGA, B. J.; HOOG, R. de; AKKERMANS, H.; VELDE, W. V. de. CommonKADS: A Comprehensive Methodology for KBS Development. **IEEE Expert**, v. 9, n. 6, p, 28-37, 1994.

SCHREIBER, G.; AKKERMANS, H.; ANJEWIERDEN, A.; HOOG, R. de; SHADBOLT, N.; VELDE, W. V. de; WIELINGA, B. **Knowkedge engineering and management**: The CommomKADS Metlhodology. 3. ed. The MIT Press, 2002.

SHADBOLT, N.; CIRAVEGNA, F.; DOMINGUE, J.; HALL, W.; MOTTA, E.; O'HARA, K.; ROBERTSON, D.; SLEEMAN, D.; TATE, A.; WILKS, Y. Advanced Knowledge Technologies at the Midterm: Tools and Methods for the Semantic Web, SHADBOLT, N.; O´HARA, K. (Ed.), Advanced Knowledge Technologies (AKT), 2004.

SILVA, F. S.da; VASCONCELOS, W.W.; ROBERTSON, D.S.; BRILHANTE, V.; MELO, A.C.V. de; FINGER, M.; AGUSTI, J. On the insufficiency of ontologies: problems in knowledge sharing and alternative solutions. **Knowledge-Based Systems**, v. 15, p. 147-167, 2002.

SMADJA, F. Retrieving collocations from Text: XTract. **Computational Linguistics**, v. 19, n. 1, p. 143-177, 1993.

SODERLAND, S.: Learning Information Extraction Rules for Semi-Structured and Free Text. **Machine Learning**, v. 34, n. 1, p. 233–272, 1999.

SOKAL, R. R.; ROHLF, F. J. The Comparison of Dendrograms by Objective Methods, **TAXON**, v. 11, p. 33-40, 1962.

TAN, A.-H. Text mining: The state of the art and the challenges. In: PACIFIC ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (PAKDD'99), WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999. **Proceedings…** 1999. p. 65-70.

TONELLA, P.; RICCA, F.; PIANTA, E.; GIRARDI, C.; DI LUCCA, G.; FASOLINO, A. R.; TRAMONTANA, P. Evaluation Methods for Web Application Clustering, In: INTERNATIONAL WORKSHOP ON WEB SITE EVOLUTION, 5., 2003. **Proceedings…** 2003.

USCHOLD, M; GRUNINGER, M. Ontologies, principles, methods and applications. **Knowledge Engineering Reviews**, v. 11, n. 2, p. 93-136, 1996.

VAN RIJSBERGEN, C. J. **Information Retrieval**. 2. ed. London: Butterworths, 1979.

VARGAS-VERA, M., MOTTA, E., DOMINGUE, J. B., LANZONI, M., STUTT, A., CIRAVEGNA, F. MnM: Ontology Driven Semi-automatic and Automatic Support for Semantic Markup. In: PROCEEDINGS OF EKAW, 2002. p. 379-391.

VECHTOMOVA, O.; ROBERTSON, S.; JONES, S. Query expansion with long-span collocates. **Information Retrieval**, v. 6, n. 2, pp. 251-273, 2003.

WANG, Y.; VECHTOMOVA, O. Exploring the Use of Term Proximity in Collocate-ranking for Query Expansion. In: JOINT ACH/ALLC (ASSOCIATION FOR COMPUTERS AND THE HUMANITIES/ASSOCIATION FOR LITERARY AND LINGUISTIC COMPUTING) CONFERENCE, 2005, Victoria, BC, Canada. **Proceedings…** Victoria, BC, Canada, 2005.

WENGER E. **Communities of practice, learning meaning and identity**. Cambridge University Press, 1998.

WILSON, T.D. The nonsense of knowledge management. **Information Research**, v. 8, n. 1, Outubro de 2002.

WITTEN, I.H.; BRAY, Z.; MAHOUI, M.; TEAHAN, B. Text mining: A new frontier for lossless compression. **Data Compression Conference**, p. 198-207, 1999.

WOHL, .D. **Intelligent text mining creates business intelligence**. 1998. Disponível em <http://www.math.tau.ac.il/~shimsh/Text_Domain/Text_Mining_IBM.pdf>. Acessado em: 10 de Abril de 2002.

YUN, C-H.; CHUANG, K-T.; CHEN, M-S. Adherence clustering: An efficient method for mining market-basket clusters. **Information Systems**, v. 31, n 3, pp. 170-186, 2006.

ZAIANE, O. R.; FOSS, A.; LEE, C-H.; WANG, W. On data clustering analysis: Scalability, constraints and validation. In: PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (PAKDD), 2002, Taipei, Taiwan. **Proceedings…** Taipei, Taiwan, 2002. p. 28-39.

ZHA, H.; DING, C.; GU, M.; HE, X.; SIMON, H. Spectral relaxation for K-means clustering. **Neural Information Processing Systems**, v. 14, p. 1057-1064, 2001.

ZHU, J.; GONÇALVES, A.; UREN, V.; MOTTA, E.; PACHECO, R. CORDER: COmmunity Relation Discovery by named entity recognition. In: INTERNATION CONFERENCE ON KNOWLEDGE CAPTURE (K-CAP 2005), 3., 2005, Banff, Canada. **Proceedings…** Banff, Canada, 2005a. p. 219-220.

ZHU, J.; UREN, V.; MOTTA, E. ESpotter: Adaptive named entity recognition for web browsing. In: WORKSHOP ON IT TOOLS FOR KNOWLEDGE MANAGEMENT SYSTEMS (WM 2005), 3., 2005, Kaiserslautern, Germany. **Proceedings…** Kaiserslautern, Germany, 2005b. p. 505-510.

Apêndice I – Tabelas de Resultado da Validação do Modelo no contexto da área de Recuperação de Informação

Todo o documento (sem janela)

k=5	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8735	0,0363	0,8014	0,0310	0,9466	0,0309	0,9404	0,0352	0,8942	0,0403	0,8748
0,2	0,0537	0,7767	0,0425	0,6480	0,0310	0,9092	0,0310	0,9106	0,0415	0,7943	0,0489	0,7767
0,3	0,0704	0,6064	0,0486	0,5154	0,0317	0,8879	0,0317	0,8855	0,0516	0,6696	0,0649	0,5956
0,4	0,0933	0,4637	0,0585	0,3857	0,0327	0,8756	0,0319	0,8688	0,0695	0,5029	0,0845	0,4474
0,5	0,1292	0,3203	0,0714	0,2933	0,0345	0,8634	0,0345	0,8643	0,0808	0,3047	0,1084	0,3119
0,6	0,1601	0,1739	0,0790	0,1947	0,0379	0,8106	0,0385	0,8424	0,1014	0,1759	0,1631	0,1556
0,7	0,2368	0,1105	0,1014	0,1292	0,0492	0,7726	0,0511	0,7347	0,1343	0,1020	0,2607	0,1010
0,8	0,3319	0,0764	0,1875	0,0788	0,0666	0,4713	0,0711	0,4989	0,2069	0,0801	0,2948	0,0604
Média	**0,1395**	**0,4252**	**0,0782**	**0,3808**	**0,0393**	**0,8171**	**0,0401**	**0,8182**	**0,0902**	**0,4405**	**0,1332**	**0,4154**

Tabela 33 - Precisão e Lembrança para expansão vetorial com k=5, sem janela, e limiar de 0.1 até 0.8

k=5	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0695	0,0600	0,0598	0,0677	0,0771
0,2	0,1005	0,0798	0,0599	0,0599	0,0790	0,0920
0,3	0,1262	0,0888	0,0612	0,0612	0,0958	0,1171
0,4	0,1554	0,1016	0,0630	0,0616	0,1220	0,1421
0,5	0,1841	0,1149	0,0664	0,0663	0,1277	0,1609
0,6	0,1667	0,1124	0,0724	0,0737	0,1286	0,1593
0,7	0,1507	0,1137	0,0925	0,0955	0,1159	0,1456
0,8	0,1242	0,1110	0,1167	0,1245	0,1155	0,1003
Média	**0,1357**	**0,0989**	**0,0740**	**0,0753**	**0,1065**	**0,1243**

Tabela 34 - Medida *F* para expansão vetorial com *k*=5, sem janela, e limiar de 0.1 até 0.8

K=10	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8755	0,0346	0,8467	0,0304	0,9592	0,0303	0,9581	0,0333	0,9320	0,0397	0,8861
0,2	0,0526	0,7833	0,0407	0,6956	0,0306	0,9393	0,0305	0,9445	0,0381	0,8422	0,0493	0,8059
0,3	0,0695	0,6190	0,0483	0,5628	0,0309	0,9187	0,0308	0,9238	0,0467	0,7568	0,0651	0,6399
0,4	0,0895	0,4750	0,0542	0,3961	0,0312	0,8973	0,0309	0,9001	0,0567	0,5894	0,0828	0,4716
0,5	0,1233	0,3382	0,0651	0,3224	0,0324	0,8524	0,0317	0,8517	0,0766	0,4193	0,1071	0,2907
0,6	0,1582	0,1801	0,0806	0,2305	0,0334	0,8160	0,0340	0,8244	0,0965	0,2524	0,1498	0,1790
0,7	0,2311	0,1162	0,1019	0,1461	0,0434	0,7963	0,0410	0,8110	0,1144	0,1261	0,2171	0,1081
0,8	0,3733	0,0639	0,1317	0,0862	0,0515	0,6822	0,0529	0,7156	0,1558	0,0763	0,3559	0,0694
Média	**0,1423**	**0,4314**	**0,0697**	**0,4108**	**0,0355**	**0,8577**	**0,0353**	**0,8661**	**0,0773**	**0,4993**	**0,1334**	**0,4313**

Tabela 35 - Precisão e Lembrança para expansão vetorial com *k*=10, sem janela, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0665	0,0590	0,0587	0,0644	0,0760
0,2	0,0986	0,0769	0,0593	0,0591	0,0728	0,0930
0,3	0,1250	0,0889	0,0599	0,0596	0,0880	0,1182
0,4	0,1506	0,0954	0,0604	0,0597	0,1035	0,1408
0,5	0,1808	0,1084	0,0625	0,0611	0,1295	0,1565
0,6	0,1684	0,1195	0,0642	0,0653	0,1396	0,1632
0,7	0,1547	0,1201	0,0823	0,0781	0,1199	0,1443
0,8	0,1091	0,1042	0,0958	0,0986	0,1024	0,1162
Média	**0,1332**	**0,0975**	**0,0679**	**0,0675**	**0,1025**	**0,1260**

Tabela 36 - Medida *F* para expansão vetorial com *k*=10, sem janela, e limiar de 0.1 até 0.8

k=15	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8767	0,0337	0,8723	0,0297	0,9762	0,0296	0,9654	0,0323	0,9435	0,0390	0,8984
0,2	0,0527	0,7930	0,0406	0,7199	0,0297	0,9671	0,0297	0,9518	0,0373	0,8906	0,0467	0,8085
0,3	0,0688	0,6059	0,0477	0,5411	0,0301	0,9594	0,0299	0,9350	0,0453	0,7897	0,0612	0,6463
0,4	0,0888	0,4792	0,0567	0,4004	0,0305	0,9518	0,0302	0,9228	0,0538	0,6373	0,0776	0,5003
0,5	0,1233	0,3586	0,0646	0,2973	0,0310	0,9256	0,0309	0,8973	0,0705	0,4482	0,0989	0,3531
0,6	0,1607	0,1973	0,0636	0,1889	0,0333	0,8829	0,0318	0,8579	0,0847	0,2639	0,1422	0,1879
0,7	0,2215	0,1190	0,0810	0,1361	0,0410	0,8691	0,0384	0,8488	0,0963	0,1449	0,1913	0,1109
0,8	0,3212	0,0779	0,1099	0,0724	0,0492	0,7638	0,0483	0,7863	0,1282	0,0639	0,2821	0,0612
Média	**0,1347**	**0,4384**	**0,0622**	**0,4035**	**0,0343**	**0,9120**	**0,0336**	**0,8957**	**0,0686**	**0,5227**	**0,1174**	**0,4458**

Tabela 37 - Precisão e Lembrança para expansão vetorial com *k*=15, sem janela, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0782	0,0648	0,0576	0,0575	0,0625	0,0747
0,2	0,0989	0,0768	0,0576	0,0576	0,0716	0,0883
0,3	0,1236	0,0876	0,0584	0,0579	0,0856	0,1118
0,4	0,1498	0,0993	0,0591	0,0585	0,0992	0,1344
0,5	0,1835	0,1062	0,0601	0,0597	0,1219	0,1545
0,6	0,1771	0,0952	0,0641	0,0613	0,1283	0,1619
0,7	0,1548	0,1015	0,0782	0,0735	0,1157	0,1404
0,8	0,1254	0,0873	0,0925	0,0910	0,0853	0,1005
Média	**0,1364**	**0,0898**	**0,0659**	**0,0646**	**0,0963**	**0,1208**

Tabela 38 - Medida *F* para expansão vetorial com *k*=15, sem janela, e limiar de 0.1 até 0.8

k=20	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8767	0,0329	0,8923	0,0293	0,9773	0,0292	0,9773	0,0314	0,9526	0,0384	0,9121
0,2	0,0524	0,7916	0,0401	0,7460	0,0293	0,9682	0,0293	0,9682	0,0360	0,9120	0,0469	0,8302
0,3	0,0693	0,6047	0,0459	0,5454	0,0297	0,9682	0,0295	0,9644	0,0427	0,8179	0,0605	0,6417
0,4	0,0888	0,4852	0,0546	0,4128	0,0299	0,9644	0,0297	0,9567	0,0494	0,6665	0,0807	0,5153
0,5	0,1257	0,3663	0,0666	0,2891	0,0303	0,9483	0,0300	0,9444	0,0670	0,4800	0,1031	0,3725
0,6	0,1732	0,2036	0,0749	0,1913	0,0315	0,9081	0,0316	0,9258	0,0821	0,2744	0,1329	0,2330
0,7	0,2162	0,1203	0,0860	0,1373	0,0369	0,9061	0,0354	0,9108	0,1111	0,1762	0,2265	0,1216
0,8	0,2991	0,0729	0,1040	0,0734	0,0436	0,8310	0,0421	0,8015	0,1162	0,0742	0,3796	0,0672
Média	**0,1332**	**0,4402**	**0,0631**	**0,4110**	**0,0326**	**0,9339**	**0,0321**	**0,9311**	**0,0670**	**0,5442**	**0,1336**	**0,4617**

Tabela 39 - Precisão e Lembrança para expansão vetorial com *k*=20, sem janela, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0635	0,0568	0,0568	0,0609	0,0736
0,2	0,0983	0,0760	0,0568	0,0569	0,0692	0,0888
0,3	0,1244	0,0847	0,0576	0,0572	0,0812	0,1105
0,4	0,1501	0,0965	0,0580	0,0576	0,0920	0,1395
0,5	0,1872	0,1083	0,0588	0,0581	0,1175	0,1614
0,6	0,1872	0,1077	0,0608	0,0611	0,1264	0,1693
0,7	0,1546	0,1057	0,0710	0,0681	0,1363	0,1582
0,8	0,1172	0,0860	0,0829	0,0800	0,0906	0,1142
Média	**0,1371**	**0,0911**	**0,0629**	**0,0620**	**0,0968**	**0,1269**

Tabela 40 - Medida *F* para expansão vetorial com *k*=20, sem janela, e limiar de 0.1 até 0.8

k=25	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8771	0,0319	0,8956	0,0292	0,9784	0,0291	0,9784	0,0309	0,9555	0,0377	0,9173
0,2	0,0519	0,7913	0,0393	0,7751	0,0292	0,9739	0,0292	0,9739	0,0345	0,9128	0,0461	0,8378
0,3	0,0697	0,6243	0,0454	0,5799	0,0294	0,9739	0,0293	0,9739	0,0403	0,8481	0,0610	0,6533
0,4	0,0879	0,4930	0,0553	0,4250	0,0296	0,9700	0,0295	0,9700	0,0471	0,7135	0,0781	0,5346
0,5	0,1274	0,3953	0,0690	0,3136	0,0299	0,9623	0,0303	0,9578	0,0658	0,5357	0,1117	0,3734
0,6	0,1757	0,2106	0,0807	0,2188	0,0311	0,9372	0,0312	0,9195	0,0826	0,3188	0,1494	0,2449
0,7	0,2147	0,1258	0,0866	0,1361	0,0347	0,9101	0,0377	0,9173	0,1073	0,1906	0,2135	0,1229
0,8	0,3212	0,0740	0,1057	0,0535	0,0405	0,8141	0,0438	0,8271	0,1470	0,0817	0,3033	0,0707
Média	**0,1362**	**0,4489**	**0,0642**	**0,4247**	**0,0317**	**0,9400**	**0,0325**	**0,9397**	**0,0694**	**0,5696**	**0,1251**	**0,4694**

Tabela 41 - Precisão e Lembrança para expansão vetorial com *k*=25, sem janela, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0617	0,0567	0,0565	0,0599	0,0724
0,2	0,0974	0,0749	0,0566	0,0567	0,0664	0,0873
0,3	0,1254	0,0842	0,0571	0,0569	0,0770	0,1115
0,4	0,1492	0,0979	0,0575	0,0572	0,0883	0,1363
0,5	0,1927	0,1131	0,0579	0,0588	0,1172	0,1720
0,6	0,1915	0,1179	0,0601	0,0604	0,1312	0,1856
0,7	0,1586	0,1059	0,0668	0,0724	0,1373	0,1560
0,8	0,1203	0,0710	0,0771	0,0833	0,1050	0,1147
Média	**0,1392**	**0,0908**	**0,0612**	**0,0628**	**0,0978**	**0,1295**

Tabela 42 - Medida *F* para expansão vetorial com *k*=25, sem janela, e limiar de 0.1 até 0.8

k=30	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0315	0,9101	0,0291	0,9822	0,0290	0,9822	0,0305	0,9597	0,0371	0,9177
0,2	0,0517	0,7910	0,0384	0,8142	0,0291	0,9777	0,0291	0,9777	0,0338	0,9373	0,0447	0,8431
0,3	0,0699	0,6265	0,0436	0,6054	0,0292	0,9777	0,0291	0,9739	0,0383	0,8730	0,0589	0,6694
0,4	0,0881	0,4956	0,0531	0,4460	0,0293	0,9700	0,0292	0,9700	0,0473	0,7308	0,0746	0,5582
0,5	0,1274	0,3987	0,0656	0,3159	0,0296	0,9655	0,0298	0,9501	0,0606	0,5805	0,1052	0,3973
0,6	0,1815	0,2139	0,0760	0,1966	0,0313	0,9503	0,0313	0,9544	0,0794	0,3713	0,1390	0,2570
0,7	0,2092	0,1237	0,0886	0,1329	0,0349	0,9161	0,0361	0,9184	0,0969	0,2060	0,2023	0,1812
0,8	0,3187	0,0771	0,0865	0,0624	0,0380	0,8144	0,0417	0,8760	0,1390	0,0994	0,3012	0,0706
Média	**0,1359**	**0,4504**	**0,0604**	**0,4354**	**0,0313**	**0,9442**	**0,0319**	**0,9503**	**0,0657**	**0,5947**	**0,1204**	**0,4868**

Tabela 43 - Precisão e Lembrança para expansão vetorial com *k*=30, sem janela, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0779	0,0608	0,0564	0,0564	0,0592	0,0713
0,2	0,0970	0,0733	0,0564	0,0565	0,0652	0,0848
0,3	0,1258	0,0814	0,0567	0,0566	0,0735	0,1082
0,4	0,1497	0,0949	0,0568	0,0567	0,0888	0,1315
0,5	0,1931	0,1087	0,0575	0,0577	0,1097	0,1664
0,6	0,1964	0,1096	0,0606	0,0607	0,1309	0,1804
0,7	0,1555	0,1063	0,0673	0,0694	0,1318	0,1912
0,8	0,1242	0,0725	0,0726	0,0797	0,1159	0,1144
Média	**0,1399**	**0,0884**	**0,0605**	**0,0617**	**0,0969**	**0,1310**

Tabela 44 - Medida *F* para expansão vetorial com *k*=30, sem janela, e limiar de 0.1 até 0.8

k=35	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0311	0,9226	0,0290	0,9833	0,0289	0,9822	0,0303	0,9665	0,0369	0,9259
0,2	0,0513	0,7943	0,0375	0,8298	0,0291	0,9822	0,0290	0,9822	0,0332	0,9422	0,0440	0,8488
0,3	0,0692	0,6321	0,0421	0,6272	0,0291	0,9777	0,0290	0,9739	0,0372	0,8887	0,0575	0,6860
0,4	0,0881	0,4987	0,0515	0,4492	0,0291	0,9700	0,0291	0,9700	0,0445	0,7521	0,0726	0,5683
0,5	0,1259	0,4002	0,0683	0,3298	0,0295	0,9578	0,0307	0,9690	0,0591	0,5866	0,1010	0,3946
0,6	0,1806	0,2209	0,0789	0,1948	0,0312	0,9509	0,0313	0,9672	0,0741	0,3823	0,1408	0,2534
0,7	0,2172	0,1242	0,0823	0,1401	0,0343	0,9160	0,0345	0,9226	0,0969	0,2300	0,2115	0,1795
0,8	0,3128	0,0801	0,1128	0,0701	0,0383	0,8584	0,0388	0,8884	0,1505	0,1068	0,2509	0,0704
Média	**0,1357**	**0,4535**	**0,0630**	**0,4455**	**0,0312**	**0,9496**	**0,0314**	**0,9570**	**0,0657**	**0,6069**	**0,1144**	**0,4909**

Tabela 45 - Precisão e Lembrança para expansão vetorial com *k*=35, sem janela, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0777	0,0602	0,0563	0,0562	0,0587	0,0710
0,2	0,0964	0,0718	0,0564	0,0563	0,0641	0,0837
0,3	0,1247	0,0788	0,0566	0,0563	0,0713	0,1061
0,4	0,1497	0,0924	0,0566	0,0565	0,0840	0,1288
0,5	0,1915	0,1131	0,0572	0,0596	0,1074	0,1609
0,6	0,1987	0,1123	0,0604	0,0607	0,1242	0,1810
0,7	0,1580	0,1037	0,0661	0,0664	0,1363	0,1942
0,8	0,1276	0,0865	0,0732	0,0744	0,1250	0,1099
Média	**0,1406**	**0,0898**	**0,0604**	**0,0608**	**0,0964**	**0,1294**

Tabela 46 - Medida *F* para expansão vetorial com *k*=35, sem janela, e limiar de 0.1 até 0.8

k=40	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8771	0,0307	0,9265	0,0289	0,9833	0,0288	0,9833	0,0300	0,9669	0,0363	0,9271
0,2	0,0517	0,8001	0,0363	0,8433	0,0289	0,9822	0,0289	0,9822	0,0323	0,9451	0,0430	0,8529
0,3	0,0688	0,6296	0,0401	0,6296	0,0291	0,9822	0,0290	0,9784	0,0356	0,8983	0,0549	0,6842
0,4	0,0877	0,5025	0,0505	0,4788	0,0291	0,9784	0,0291	0,9784	0,0420	0,7778	0,0694	0,5815
0,5	0,1231	0,4030	0,0674	0,3494	0,0293	0,9495	0,0306	0,9779	0,0544	0,6118	0,0932	0,4131
0,6	0,1692	0,2628	0,0813	0,2164	0,0313	0,9790	0,0309	0,9733	0,0713	0,4095	0,1379	0,2700
0,7	0,2187	0,1302	0,0991	0,1487	0,0336	0,9298	0,0333	0,9291	0,0941	0,2578	0,1928	0,1831
0,8	0,2977	0,0776	0,1312	0,0702	0,0361	0,8259	0,0349	0,8455	0,1422	0,1209	0,2300	0,0629
Média	**0,1322**	**0,4604**	**0,0671**	**0,4579**	**0,0308**	**0,9513**	**0,0307**	**0,9560**	**0,0627**	**0,6235**	**0,1072**	**0,4968**

Tabela 47 - Precisão e Lembrança para expansão vetorial com *k*=40, sem janela, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0594	0,0561	0,0560	0,0581	0,0699
0,2	0,0972	0,0696	0,0562	0,0561	0,0625	0,0818
0,3	0,1241	0,0754	0,0564	0,0563	0,0684	0,1017
0,4	0,1494	0,0913	0,0565	0,0566	0,0797	0,1240
0,5	0,1886	0,1130	0,0568	0,0593	0,1000	0,1521
0,6	0,2058	0,1182	0,0607	0,0600	0,1214	0,1826
0,7	0,1632	0,1190	0,0649	0,0643	0,1379	0,1878
0,8	0,1232	0,0915	0,0691	0,0670	0,1307	0,0987
Média	**0,1411**	**0,0922**	**0,0596**	**0,0595**	**0,0948**	**0,1248**

Tabela 48 - Medida *F* para expansão vetorial com *k*=40, sem janela, e limiar de 0.1 até 0.8

k=45	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8782	0,0305	0,9324	0,0288	0,9833	0,0288	0,9833	0,0299	0,9784	0,0357	0,9275
0,2	0,0515	0,8057	0,0357	0,8568	0,0289	0,9822	0,0288	0,9822	0,0317	0,9502	0,0418	0,8548
0,3	0,0684	0,6276	0,0393	0,6592	0,0290	0,9784	0,0289	0,9784	0,0347	0,9162	0,0541	0,7069
0,4	0,0862	0,5014	0,0492	0,5028	0,0290	0,9784	0,0291	0,9784	0,0402	0,8001	0,0682	0,5937
0,5	0,1221	0,4030	0,0611	0,3513	0,0296	0,9739	0,0306	0,9900	0,0511	0,6830	0,0902	0,4290
0,6	0,1674	0,2654	0,0770	0,2306	0,0310	0,9877	0,0308	0,9871	0,0717	0,4771	0,1337	0,2836
0,7	0,2289	0,1317	0,1026	0,1520	0,0325	0,9435	0,0325	0,9462	0,0958	0,2780	0,1822	0,1908
0,8	0,3062	0,0884	0,1434	0,0942	0,0350	0,8636	0,0343	0,8710	0,1433	0,1336	0,2279	0,0578
Média	**0,1339**	**0,4627**	**0,0673**	**0,4724**	**0,0305**	**0,9614**	**0,0305**	**0,9646**	**0,0623**	**0,6521**	**0,1042**	**0,5055**

Tabela 49 - Precisão e Lembrança para expansão vetorial com *k*=45, sem janela, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0591	0,0560	0,0559	0,0580	0,0687
0,2	0,0968	0,0685	0,0561	0,0560	0,0614	0,0798
0,3	0,1233	0,0742	0,0563	0,0562	0,0669	0,1006
0,4	0,1471	0,0895	0,0564	0,0564	0,0766	0,1223
0,5	0,1874	0,1041	0,0574	0,0594	0,0950	0,1491
0,6	0,2053	0,1155	0,0601	0,0598	0,1247	0,1818
0,7	0,1672	0,1225	0,0629	0,0629	0,1425	0,1864
0,8	0,1372	0,1137	0,0673	0,0660	0,1383	0,0922
Média	**0,1427**	**0,0934**	**0,0591**	**0,0591**	**0,0954**	**0,1226**

Tabela 50 - Medida *F* para expansão vetorial com *k*=45, sem janela, e limiar de 0.1 até 0.8

k=50	LRD		Phi-squared		MI		VMI		Z score		LSI	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0405	0,8782	0,0302	0,9307	0,0288	0,9833	0,0287	0,9833	0,0298	0,9951	0,0351	0,9344
0,2	0,0517	0,8117	0,0356	0,8659	0,0288	0,9833	0,0288	0,9833	0,0314	0,9681	0,0411	0,8573
0,3	0,0681	0,6315	0,0404	0,6954	0,0289	0,9795	0,0289	0,9795	0,0341	0,9251	0,0524	0,7228
0,4	0,0871	0,5098	0,0505	0,5293	0,0290	0,9795	0,0290	0,9795	0,0397	0,8264	0,0668	0,6059
0,5	0,1155	0,3898	0,0719	0,3720	0,0305	0,9960	0,0305	0,9960	0,0535	0,7200	0,0892	0,4528
0,6	0,1565	0,2549	0,0848	0,2260	0,0307	0,9889	0,0305	0,9937	0,0680	0,5162	0,1286	0,2867
0,7	0,2240	0,1363	0,0932	0,1373	0,0312	0,9528	0,0319	0,9619	0,0918	0,3021	0,1797	0,1910
0,8	0,2991	0,0889	0,1100	0,0737	0,0328	0,8696	0,0338	0,8871	0,1380	0,1422	0,2223	0,0649
Média	**0,1303**	**0,4626**	**0,0646**	**0,4788**	**0,0301**	**0,9666**	**0,0303**	**0,9705**	**0,0608**	**0,6744**	**0,1019**	**0,5145**

Tabela 51 - Precisão e Lembrança para expansão vetorial com *k*=50, sem janela, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score	LSI
Limiar	**F**	**F**	**F**	**F**	**F**	**F**
0,1	0,0774	0,0586	0,0559	0,0559	0,0580	0,0676
0,2	0,0973	0,0684	0,0560	0,0559	0,0609	0,0784
0,3	0,1229	0,0764	0,0562	0,0562	0,0658	0,0978
0,4	0,1488	0,0922	0,0562	0,0563	0,0757	0,1204
0,5	0,1782	0,1205	0,0591	0,0591	0,0996	0,1490
0,6	0,1939	0,1233	0,0596	0,0592	0,1201	0,1775
0,7	0,1695	0,1110	0,0605	0,0617	0,1408	0,1852
0,8	0,1370	0,0883	0,0631	0,0650	0,1401	0,1005
Média	**0,1406**	**0,0923**	**0,0583**	**0,0587**	**0,0951**	**0,1221**

Tabela 52 - Medida *F* para expansão vetorial com *k*=50, sem janela, e limiar de 0.1 até 0.8

Janela de 20

k=5	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8735	0,0448	0,7147	0,0310	0,9466	0,0310	0,9426	0,0418	0,8656
0,2	0,0536	0,7772	0,0527	0,4528	0,0310	0,9154	0,0312	0,9090	0,0514	0,7229
0,3	0,0706	0,6085	0,0641	0,3170	0,0316	0,8941	0,0319	0,8883	0,0652	0,5663
0,4	0,0931	0,4586	0,0815	0,2276	0,0326	0,8819	0,0310	0,8573	0,0852	0,3680
0,5	0,1285	0,3190	0,1045	0,1682	0,0344	0,8634	0,0349	0,8702	0,1072	0,2153
0,6	0,1589	0,1705	0,1132	0,1039	0,0382	0,8070	0,0385	0,8321	0,1334	0,1297
0,7	0,2294	0,1105	0,2131	0,0716	0,0474	0,7402	0,0522	0,7227	0,1946	0,0933
0,8	0,3692	0,0677	0,1802	0,0522	0,0659	0,5186	0,0761	0,4908	0,2695	0,0409
Média	**0,1430**	**0,4232**	**0,1068**	**0,2635**	**0,0390**	**0,8209**	**0,0409**	**0,8141**	**0,1185**	**0,3753**

Tabela 53 - Precisão e Lembrança para expansão vetorial com *k*=5, janela de 20, e limiar de 0.1 até 0.8

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0843	0,0599	0,0600	0,0798
0,2	0,1003	0,0943	0,0600	0,0603	0,0960
0,3	0,1265	0,1066	0,0611	0,0617	0,1169
0,4	0,1548	0,1200	0,0630	0,0599	0,1383
0,5	0,1832	0,1290	0,0662	0,0671	0,1431
0,6	0,1645	0,1083	0,0729	0,0736	0,1315
0,7	0,1492	0,1072	0,0891	0,0974	0,1261
0,8	0,1145	0,0809	0,1169	0,1318	0,0710
Média	**0,1339**	**0,1038**	**0,0736**	**0,0765**	**0,1129**

Tabela 54 - Medida *F* para expansão vetorial com *k*=5, janela de 20, e limiar de 0.1 até 0.8

k=10	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8755	0,0427	0,7327	0,0305	0,9643	0,0303	0,9581	0,0423	0,8629
0,2	0,0526	0,7894	0,0514	0,5288	0,0308	0,9408	0,0307	0,9406	0,0519	0,7133
0,3	0,0689	0,6188	0,0623	0,3662	0,0310	0,9195	0,0310	0,9238	0,0688	0,5579
0,4	0,0889	0,4777	0,0803	0,2851	0,0311	0,8912	0,0308	0,8886	0,0918	0,3667
0,5	0,1276	0,3504	0,0978	0,1968	0,0322	0,8477	0,0320	0,8479	0,1072	0,2196
0,6	0,1726	0,1917	0,1117	0,1288	0,0359	0,8184	0,0341	0,8141	0,1368	0,1359
0,7	0,2213	0,1150	0,1400	0,0740	0,0440	0,8300	0,0431	0,8447	0,2258	0,0935
0,8	0,3507	0,0709	0,1578	0,0472	0,0532	0,6749	0,0535	0,7030	0,3082	0,0532
Média	**0,1404**	**0,4362**	**0,0930**	**0,2950**	**0,0361**	**0,8609**	**0,0357**	**0,8651**	**0,1291**	**0,3754**

Tabela 55 - Precisão e Lembrança para expansão vetorial com *k*=10, janela de 20, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0808	0,0591	0,0588	0,0807
0,2	0,0986	0,0937	0,0596	0,0594	0,0967
0,3	0,1239	0,1065	0,0600	0,0599	0,1225
0,4	0,1499	0,1253	0,0601	0,0595	0,1469
0,5	0,1871	0,1307	0,0620	0,0617	0,1440
0,6	0,1816	0,1196	0,0688	0,0655	0,1363
0,7	0,1513	0,0968	0,0836	0,0820	0,1322
0,8	0,1179	0,0727	0,0987	0,0994	0,0907
Avg	**0,1361**	**0,1033**	**0,0690**	**0,0683**	**0,1188**

Tabela 56 - Medida *F* para expansão vetorial com *k*=10, janela de 20, e limiar de 0.1 até 0.8

k=15	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8761	0,0415	0,7553	0,0297	0,9751	0,0297	0,9706	0,0422	0,8620
0,2	0,0524	0,7979	0,0514	0,5684	0,0297	0,9660	0,0298	0,9493	0,0526	0,7037
0,3	0,0697	0,6340	0,0610	0,3939	0,0302	0,9622	0,0300	0,9402	0,0693	0,5291
0,4	0,0870	0,4840	0,0825	0,2981	0,0305	0,9461	0,0300	0,9126	0,0968	0,3385
0,5	0,1308	0,3760	0,0916	0,1788	0,0307	0,9060	0,0307	0,8885	0,1298	0,2199
0,6	0,1715	0,2022	0,1191	0,1252	0,0326	0,8665	0,0323	0,8555	0,1700	0,1340
0,7	0,2160	0,1199	0,1478	0,0653	0,0420	0,8835	0,0398	0,8726	0,2296	0,0695
0,8	0,3182	0,0766	0,1409	0,0418	0,0479	0,7380	0,0483	0,7730	0,3344	0,0461
Média	**0,1358**	**0,4458**	**0,0920**	**0,3034**	**0,0342**	**0,9054**	**0,0338**	**0,8953**	**0,1406**	**0,3629**

Tabela 57 - Precisão e Lembrança para expansão vetorial com *k*=15, janela de 20, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0787	0,0576	0,0576	0,0805
0,2	0,0983	0,0942	0,0577	0,0577	0,0979
0,3	0,1256	0,1056	0,0586	0,0582	0,1225
0,4	0,1475	0,1292	0,0591	0,0581	0,1505
0,5	0,1940	0,1212	0,0594	0,0593	0,1632
0,6	0,1856	0,1221	0,0628	0,0622	0,1499
0,7	0,1542	0,0906	0,0802	0,0761	0,1068
0,8	0,1234	0,0645	0,0899	0,0909	0,0810
Avg	**0,1383**	**0,1008**	**0,0656**	**0,0650**	**0,1191**

Tabela 58 - Medida *F* para expansão vetorial com *k*=15, janela de 20, e limiar de 0.1 até 0.8

k=20	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8767	0,0408	0,7955	0,0293	0,9773	0,0293	0,9773	0,0423	0,8652
0,2	0,0518	0,7979	0,0494	0,5704	0,0293	0,9682	0,0294	0,9682	0,0532	0,6945
0,3	0,0690	0,6329	0,0586	0,3956	0,0297	0,9682	0,0296	0,9682	0,0699	0,5253
0,4	0,0886	0,4934	0,0813	0,2880	0,0300	0,9543	0,0298	0,9605	0,0979	0,3385
0,5	0,1258	0,3814	0,0818	0,1544	0,0305	0,9382	0,0304	0,9382	0,1158	0,2103
0,6	0,1658	0,2043	0,1061	0,1004	0,0315	0,8952	0,0310	0,8980	0,1344	0,1146
0,7	0,2168	0,1284	0,1336	0,0674	0,0372	0,8887	0,0370	0,9020	0,2503	0,0738
0,8	0,3095	0,0801	0,1842	0,0468	0,0439	0,8137	0,0446	0,8514	0,3619	0,0411
Média	**0,1335**	**0,4494**	**0,0920**	**0,3023**	**0,0327**	**0,9255**	**0,0326**	**0,9330**	**0,1407**	**0,3579**

Tabela 59 - Precisão e Lembrança para expansão vetorial com *k*=20, janela de 20, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0775	0,0570	0,0569	0,0807
0,2	0,0974	0,0910	0,0569	0,0570	0,0988
0,3	0,1245	0,1021	0,0577	0,0574	0,1234
0,4	0,1502	0,1268	0,0581	0,0578	0,1519
0,5	0,1892	0,1069	0,0591	0,0589	0,1494
0,6	0,1831	0,1032	0,0608	0,0599	0,1237
0,7	0,1613	0,0896	0,0714	0,0710	0,1140
0,8	0,1272	0,0746	0,0833	0,0847	0,0738
Média	**0,1388**	**0,0965**	**0,0630**	**0,0630**	**0,1145**

Tabela 60 - Medida *F* para expansão vetorial com *k*=20, janela de 20, e limiar de 0.1 até 0.8

k=25	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8767	0,0398	0,8151	0,0292	0,9784	0,0291	0,9784	0,0419	0,8647
0,2	0,0517	0,7952	0,0490	0,6040	0,0292	0,9693	0,0292	0,9693	0,0529	0,6831
0,3	0,0682	0,6116	0,0582	0,4147	0,0295	0,9693	0,0292	0,9693	0,0697	0,5200
0,4	0,0893	0,5046	0,0770	0,2905	0,0297	0,9616	0,0293	0,9655	0,0947	0,3242
0,5	0,1246	0,3979	0,0860	0,1594	0,0302	0,9393	0,0301	0,9455	0,1169	0,2060
0,6	0,1724	0,2156	0,1117	0,0979	0,0310	0,9150	0,0320	0,9507	0,1548	0,1393
0,7	0,2135	0,1271	0,1376	0,0645	0,0371	0,8942	0,0374	0,9187	0,2340	0,0733
0,8	0,2974	0,0903	0,1617	0,0434	0,0435	0,8492	0,0430	0,8188	0,2913	0,0370
Média	**0,1322**	**0,4524**	**0,0901**	**0,3112**	**0,0324**	**0,9345**	**0,0324**	**0,9395**	**0,1320**	**0,3560**

Tabela 61 - Precisão e Lembrança para expansão vetorial com *k*=25, janela de 20, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0779	0,0758	0,0568	0,0566	0,0799
0,2	0,0971	0,0906	0,0567	0,0566	0,0983
0,3	0,1226	0,1022	0,0572	0,0568	0,1230
0,4	0,1518	0,1218	0,0575	0,0570	0,1466
0,5	0,1898	0,1117	0,0586	0,0583	0,1492
0,6	0,1916	0,1043	0,0601	0,0619	0,1467
0,7	0,1594	0,0878	0,0712	0,0718	0,1116
0,8	0,1386	0,0684	0,0827	0,0817	0,0656
Média	**0,1411**	**0,0953**	**0,0626**	**0,0626**	**0,1151**

Tabela 62 - Medida F para expansão vetorial com *k*=25, janela de 20, e limiar de 0.1 até 0.8

k=30	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8806	0,0384	0,8314	0,0291	0,9822	0,0290	0,9822	0,0417	0,8690
0,2	0,0518	0,8085	0,0475	0,6168	0,0291	0,9777	0,0291	0,9777	0,0516	0,6931
0,3	0,0680	0,6159	0,0578	0,4269	0,0293	0,9777	0,0291	0,9739	0,0676	0,5183
0,4	0,0887	0,5073	0,0720	0,2922	0,0293	0,9700	0,0292	0,9700	0,0843	0,3839
0,5	0,1240	0,4094	0,0908	0,1765	0,0299	0,9578	0,0299	0,9539	0,1104	0,2097
0,6	0,1690	0,2187	0,0940	0,1009	0,0317	0,9539	0,0318	0,9625	0,1369	0,1498
0,7	0,2163	0,1296	0,1101	0,0620	0,0359	0,9113	0,0353	0,9055	0,1752	0,0799
0,8	0,2936	0,0903	0,1539	0,0400	0,0410	0,8554	0,0412	0,8644	0,2552	0,0449
Média	**0,1315**	**0,4575**	**0,0831**	**0,3183**	**0,0319**	**0,9482**	**0,0318**	**0,9488**	**0,1154**	**0,3686**

Tabela 63 - Precisão e Lembrança para expansão vetorial com *k*=30, janela de 20, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0779	0,0734	0,0565	0,0564	0,0795
0,2	0,0974	0,0883	0,0565	0,0565	0,0961
0,3	0,1224	0,1018	0,0568	0,0566	0,1197
0,4	0,1509	0,1156	0,0568	0,0567	0,1382
0,5	0,1904	0,1199	0,0580	0,0579	0,1447
0,6	0,1906	0,0974	0,0613	0,0615	0,1430
0,7	0,1621	0,0793	0,0691	0,0680	0,1098
0,8	0,1382	0,0635	0,0783	0,0786	0,0763
Média	**0,1412**	**0,0924**	**0,0617**	**0,0615**	**0,1134**

Tabela 64 - Medida *F* para expansão vetorial com *k*=30, janela de 20, e limiar de 0.1 até 0.8

k=35	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8806	0,0372	0,8475	0,0290	0,9833	0,0289	0,9822	0,0411	0,8736
0,2	0,0515	0,8043	0,0469	0,6429	0,0291	0,9822	0,0289	0,9822	0,0507	0,7109
0,3	0,0681	0,6204	0,0582	0,4601	0,0291	0,9777	0,0289	0,9739	0,0648	0,5579
0,4	0,0884	0,5098	0,0736	0,2943	0,0292	0,9700	0,0291	0,9700	0,0834	0,4098
0,5	0,1216	0,4103	0,0878	0,1791	0,0299	0,9578	0,0309	0,9690	0,1129	0,2423
0,6	0,1616	0,2694	0,1017	0,0982	0,0323	0,9479	0,0323	0,9588	0,1304	0,1652
0,7	0,2247	0,1332	0,0919	0,0476	0,0340	0,9040	0,0337	0,9078	0,1627	0,1004
0,8	0,2901	0,0913	0,1627	0,0365	0,0404	0,8537	0,0392	0,8283	0,2128	0,0558
Média	**0,1308**	**0,4649**	**0,0825**	**0,3258**	**0,0316**	**0,9471**	**0,0315**	**0,9465**	**0,1074**	**0,3895**

Tabela 65 - Precisão e Lembrança para expansão vetorial com *k*=35, janela de 20, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0778	0,0713	0,0563	0,0561	0,0786
0,2	0,0968	0,0874	0,0564	0,0562	0,0946
0,3	0,1227	0,1034	0,0566	0,0562	0,1161
0,4	0,1507	0,1178	0,0566	0,0565	0,1386
0,5	0,1877	0,1179	0,0579	0,0600	0,1540
0,6	0,2020	0,0999	0,0624	0,0625	0,1458
0,7	0,1672	0,0627	0,0656	0,0650	0,1241
0,8	0,1389	0,0596	0,0771	0,0749	0,0885
Média	**0,1430**	**0,0900**	**0,0611**	**0,0609**	**0,1175**

Tabela 66 - Medida *F* para expansão vetorial com *k*=35, janela de 20, e limiar de 0.1 até 0.8

k=40	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8809	0,0364	0,8657	0,0288	0,9833	0,0288	0,9833	0,0406	0,8767
0,2	0,0515	0,8076	0,0460	0,6653	0,0289	0,9822	0,0289	0,9822	0,0499	0,7223
0,3	0,0681	0,6197	0,0577	0,4819	0,0290	0,9777	0,0289	0,9739	0,0634	0,5614
0,4	0,0894	0,5174	0,0710	0,3174	0,0290	0,9700	0,0290	0,9700	0,0838	0,4288
0,5	0,1214	0,4106	0,0869	0,1795	0,0310	0,9812	0,0307	0,9812	0,1114	0,2598
0,6	0,1624	0,2755	0,0962	0,0970	0,0319	0,9707	0,0314	0,9641	0,1271	0,1739
0,7	0,2284	0,1346	0,1041	0,0524	0,0325	0,9115	0,0325	0,9225	0,1647	0,1040
0,8	0,2951	0,0912	0,1341	0,0327	0,0375	0,8295	0,0357	0,8486	0,2252	0,0524
Média	**0,1321**	**0,4672**	**0,0790**	**0,3365**	**0,0311**	**0,9508**	**0,0307**	**0,9532**	**0,1083**	**0,3974**

Tabela 67 - Precisão e Lembrança para expansão vetorial com *k*=40, janela de 20, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0778	0,0699	0,0560	0,0560	0,0777
0,2	0,0968	0,0860	0,0562	0,0561	0,0934
0,3	0,1228	0,1030	0,0564	0,0561	0,1139
0,4	0,1524	0,1161	0,0564	0,0563	0,1402
0,5	0,1874	0,1171	0,0600	0,0596	0,1559
0,6	0,2043	0,0966	0,0618	0,0607	0,1469
0,7	0,1694	0,0697	0,0627	0,0628	0,1275
0,8	0,1394	0,0526	0,0717	0,0685	0,0850
Média	**0,1438**	**0,0889**	**0,0602**	**0,0595**	**0,1176**

Tabela 68 - Medida *F* para expansão vetorial com *k*=40, janela de 20, e limiar de 0.1 até 0.8

k=45	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8828	0,0359	0,8727	0,0288	0,9833	0,0288	0,9833	0,0404	0,8784
0,2	0,0518	0,8103	0,0454	0,6816	0,0289	0,9822	0,0288	0,9822	0,0493	0,7333
0,3	0,0682	0,6197	0,0570	0,4885	0,0289	0,9777	0,0288	0,9739	0,0641	0,5863
0,4	0,0892	0,5192	0,0684	0,3047	0,0290	0,9700	0,0290	0,9739	0,0829	0,4435
0,5	0,1228	0,4189	0,0893	0,1796	0,0307	0,9812	0,0306	0,9852	0,1096	0,2753
0,6	0,1623	0,2753	0,1015	0,0990	0,0312	0,9730	0,0311	0,9716	0,1249	0,1703
0,7	0,2299	0,1395	0,0990	0,0566	0,0319	0,9307	0,0319	0,9294	0,1533	0,1046
0,8	0,3088	0,0809	0,1276	0,0320	0,0338	0,8501	0,0338	0,8507	0,2353	0,0584
Média	**0,1342**	**0,4683**	**0,0780**	**0,3393**	**0,0304**	**0,9560**	**0,0304**	**0,9563**	**0,1075**	**0,4063**

Tabela 69 - Precisão e Lembrança para expansão vetorial com *k*=45, janela de 20, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0777	0,0690	0,0560	0,0559	0,0773
0,2	0,0973	0,0851	0,0561	0,0560	0,0924
0,3	0,1229	0,1021	0,0562	0,0560	0,1155
0,4	0,1522	0,1117	0,0563	0,0563	0,1397
0,5	0,1899	0,1193	0,0595	0,0594	0,1568
0,6	0,2042	0,1003	0,0605	0,0602	0,1441
0,7	0,1736	0,0720	0,0617	0,0617	0,1244
0,8	0,1282	0,0511	0,0651	0,0650	0,0936
Média	**0,1433**	**0,0888**	**0,0589**	**0,0588**	**0,1180**

Tabela 70 - Medida *F* para expansão vetorial com *k*=45, janela de 20, e limiar de 0.1 até 0.8

k=50	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8874	0,0353	0,8812	0,0288	0,9833	0,0287	0,9833	0,0399	0,8856
0,2	0,0516	0,8029	0,0446	0,6995	0,0288	0,9822	0,0288	0,9822	0,0489	0,7492
0,3	0,0684	0,6161	0,0540	0,5042	0,0289	0,9784	0,0288	0,9784	0,0627	0,5975
0,4	0,0888	0,5185	0,0706	0,3217	0,0290	0,9739	0,0289	0,9739	0,0833	0,4697
0,5	0,1239	0,4122	0,0830	0,1688	0,0305	0,9900	0,0304	0,9900	0,1037	0,2943
0,6	0,1655	0,2698	0,0887	0,0895	0,0307	0,9793	0,0306	0,9782	0,1235	0,1848
0,7	0,2257	0,1341	0,1217	0,0599	0,0309	0,9330	0,0317	0,9376	0,1644	0,1250
0,8	0,2974	0,0776	0,1293	0,0314	0,0329	0,8632	0,0340	0,8641	0,2112	0,0669
Média	**0,1327**	**0,4648**	**0,0784**	**0,3445**	**0,0300**	**0,9604**	**0,0302**	**0,9610**	**0,1047**	**0,4216**

Tabela 71 - Precisão e Lembrança para expansão vetorial com *k*=50, janela de 20, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0778	0,0678	0,0559	0,0559	0,0764
0,2	0,0970	0,0838	0,0560	0,0559	0,0918
0,3	0,1232	0,0975	0,0562	0,0560	0,1134
0,4	0,1516	0,1158	0,0562	0,0561	0,1415
0,5	0,1905	0,1113	0,0591	0,0590	0,1534
0,6	0,2052	0,0891	0,0595	0,0593	0,1481
0,7	0,1682	0,0802	0,0598	0,0614	0,1420
0,8	0,1231	0,0505	0,0633	0,0655	0,1016
Média	**0,1421**	**0,0870**	**0,0583**	**0,0586**	**0,1210**

Tabela 72 - Medida *F* para expansão vetorial com *k*=50, janela de 20, e limiar de 0.1 até 0.8

Janela de 50

k=5	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8735	0,0316	0,8347	0,0309	0,9404	0,0309	0,9404	0,0416	0,8659
0,2	0,0536	0,7725	0,0347	0,6716	0,0309	0,9029	0,0310	0,9106	0,0514	0,7512
0,3	0,0704	0,6085	0,0392	0,5403	0,0316	0,8816	0,0317	0,8855	0,0639	0,5751
0,4	0,0927	0,4627	0,0458	0,4026	0,0326	0,8694	0,0320	0,8688	0,0839	0,3916
0,5	0,1285	0,3231	0,0639	0,3166	0,0349	0,8605	0,0346	0,8672	0,1036	0,2423
0,6	0,1563	0,1694	0,0790	0,2051	0,0385	0,8065	0,0389	0,8445	0,1323	0,1561
0,7	0,2327	0,1070	0,1183	0,1212	0,0502	0,7707	0,0519	0,7299	0,1637	0,0999
0,8	0,3262	0,0659	0,1051	0,0683	0,0707	0,4884	0,0707	0,4939	0,2817	0,0623
Média	**0,1377**	**0,4228**	**0,0647**	**0,3951**	**0,0400**	**0,8150**	**0,0402**	**0,8176**	**0,1153**	**0,3931**

Tabela 73 - Precisão e Lembrança para expansão vetorial com *k*=5, janela de 50, e limiar de 0.1 até 0.8

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0608	0,0599	0,0598	0,0794
0,2	0,1003	0,0661	0,0597	0,0600	0,0963
0,3	0,1262	0,0731	0,0611	0,0611	0,1150
0,4	0,1545	0,0823	0,0629	0,0617	0,1382
0,5	0,1838	0,1063	0,0671	0,0665	0,1451
0,6	0,1626	0,1140	0,0735	0,0743	0,1432
0,7	0,1466	0,1197	0,0942	0,0970	0,1241
0,8	0,1096	0,0828	0,1235	0,1237	0,1020
Média	**0,1327**	**0,0881**	**0,0752**	**0,0755**	**0,1179**

Tabela 74 - Medida *F* para expansão vetorial com *k*=5, janela de 50, e limiar de 0.1 até 0.8

k=10	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8761	0,0317	0,8855	0,0303	0,9581	0,0303	0,9518	0,0418	0,8683
0,2	0,0527	0,7850	0,0356	0,7332	0,0306	0,9445	0,0304	0,9320	0,0516	0,7435
0,3	0,0690	0,6213	0,0394	0,5859	0,0309	0,9238	0,0306	0,9113	0,0645	0,5731
0,4	0,0897	0,4766	0,0463	0,4315	0,0313	0,8900	0,0307	0,8938	0,0853	0,3875
0,5	0,1235	0,3430	0,0569	0,3294	0,0319	0,8451	0,0316	0,8517	0,1018	0,2248
0,6	0,1604	0,1851	0,0746	0,2334	0,0334	0,8145	0,0344	0,8212	0,1451	0,1577
0,7	0,2225	0,1135	0,0880	0,1459	0,0429	0,7881	0,0420	0,8130	0,2091	0,1052
0,8	0,3339	0,0667	0,1075	0,0730	0,0549	0,6691	0,0546	0,7489	0,3097	0,0567
Média	**0,1366**	**0,4334**	**0,0600**	**0,4272**	**0,0358**	**0,8541**	**0,0356**	**0,8655**	**0,1261**	**0,3896**

Tabela 75 - Precisão e Lembrança para expansão vetorial com *k*=10, janela de 50, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0782	0,0612	0,0588	0,0586	0,0797
0,2	0,0988	0,0679	0,0592	0,0589	0,0965
0,3	0,1242	0,0738	0,0599	0,0592	0,1160
0,4	0,1510	0,0836	0,0605	0,0593	0,1398
0,5	0,1816	0,0970	0,0615	0,0609	0,1402
0,6	0,1718	0,1130	0,0642	0,0661	0,1511
0,7	0,1503	0,1098	0,0815	0,0798	0,1400
0,8	0,1112	0,0869	0,1015	0,1019	0,0959
Média	**0,1334**	**0,0866**	**0,0684**	**0,0681**	**0,1199**

Tabela 76 - Medida *F* para expansão vetorial com *k*=10, janela de 50, e limiar de 0.1 até 0.8

k=15	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8767	0,0310	0,8943	0,0296	0,9762	0,0297	0,9654	0,0416	0,8675
0,2	0,0526	0,7931	0,0351	0,7474	0,0297	0,9671	0,0297	0,9518	0,0515	0,7386
0,3	0,0694	0,6089	0,0387	0,5765	0,0302	0,9633	0,0299	0,9350	0,0635	0,5639
0,4	0,0888	0,4791	0,0425	0,4238	0,0307	0,9556	0,0302	0,9228	0,0852	0,3861
0,5	0,1258	0,3642	0,0509	0,2976	0,0315	0,9256	0,0308	0,8973	0,1095	0,2480
0,6	0,1673	0,2005	0,0709	0,2194	0,0335	0,8886	0,0316	0,8585	0,1379	0,1579
0,7	0,2251	0,1192	0,0736	0,1043	0,0427	0,9094	0,0379	0,8510	0,1812	0,0873
0,8	0,3464	0,0815	0,1125	0,0532	0,0485	0,7550	0,0478	0,7726	0,2738	0,0587
Média	**0,1395**	**0,4404**	**0,0569**	**0,4146**	**0,0346**	**0,9176**	**0,0335**	**0,8943**	**0,1180**	**0,3885**

Tabela 77 - Precisão e Lembrança para expansão vetorial com *k*=15, janela de 50, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0782	0,0600	0,0575	0,0575	0,0794
0,2	0,0986	0,0671	0,0576	0,0576	0,0963
0,3	0,1247	0,0725	0,0586	0,0579	0,1142
0,4	0,1499	0,0772	0,0595	0,0585	0,1396
0,5	0,1870	0,0870	0,0609	0,0596	0,1520
0,6	0,1824	0,1071	0,0645	0,0610	0,1473
0,7	0,1558	0,0863	0,0815	0,0726	0,1178
0,8	0,1319	0,0722	0,0912	0,0900	0,0967
Média	**0,1386**	**0,0787**	**0,0664**	**0,0644**	**0,1179**

Tabela 78 - Medida *F* para expansão vetorial com *k*=15, janela de 50, e limiar de 0.1 até 0.8

k=20	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8767	0,0309	0,9034	0,0293	0,9773	0,0293	0,9784	0,0417	0,8692
0,2	0,0519	0,7887	0,0357	0,7805	0,0293	0,9682	0,0293	0,9693	0,0519	0,7162
0,3	0,0687	0,6125	0,0389	0,6033	0,0297	0,9682	0,0295	0,9655	0,0653	0,5638
0,4	0,0879	0,4915	0,0422	0,4247	0,0298	0,9605	0,0297	0,9578	0,0860	0,3794
0,5	0,1243	0,3869	0,0559	0,3270	0,0302	0,9483	0,0300	0,9455	0,1127	0,2518
0,6	0,1645	0,2136	0,0691	0,2294	0,0314	0,9111	0,0315	0,9269	0,1469	0,1540
0,7	0,2128	0,1247	0,0794	0,1300	0,0364	0,9084	0,0356	0,9151	0,2253	0,0845
0,8	0,3210	0,0912	0,1019	0,0539	0,0439	0,7877	0,0425	0,8040	0,3392	0,0442
Média	**0,1340**	**0,4482**	**0,0568**	**0,4315**	**0,0325**	**0,9287**	**0,0322**	**0,9328**	**0,1336**	**0,3829**

Tabela 79 - Precisão e Lembrança para expansão vetorial com *k*=20, janela de 50, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0598	0,0568	0,0569	0,0796
0,2	0,0974	0,0683	0,0568	0,0569	0,0969
0,3	0,1235	0,0731	0,0576	0,0572	0,1171
0,4	0,1491	0,0767	0,0579	0,0576	0,1402
0,5	0,1881	0,0955	0,0585	0,0581	0,1557
0,6	0,1859	0,1062	0,0607	0,0610	0,1504
0,7	0,1572	0,0986	0,0699	0,0686	0,1230
0,8	0,1421	0,0705	0,0833	0,0807	0,0783
Média	**0,1402**	**0,0811**	**0,0627**	**0,0621**	**0,1176**

Tabela 80 - Medida *F* para expansão vetorial com *k*=20, janela de 50, e limiar de 0.1 até 0.8

k=25	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8771	0,0301	0,9149	0,0292	0,9784	0,0291	0,9784	0,0417	0,8698
0,2	0,0517	0,7914	0,0338	0,8014	0,0292	0,9739	0,0292	0,9739	0,0529	0,7229
0,3	0,0688	0,6134	0,0382	0,6158	0,0294	0,9739	0,0293	0,9739	0,0666	0,5500
0,4	0,0883	0,4984	0,0447	0,4271	0,0296	0,9700	0,0295	0,9700	0,0946	0,3829
0,5	0,1241	0,3979	0,0630	0,2599	0,0302	0,9616	0,0304	0,9578	0,1138	0,2387
0,6	0,1744	0,2281	0,0948	0,1692	0,0312	0,9361	0,0310	0,9160	0,1665	0,1524
0,7	0,2164	0,1255	0,1096	0,0897	0,0365	0,9122	0,0376	0,9184	0,2335	0,0825
0,8	0,3201	0,0912	0,1605	0,0414	0,0413	0,8211	0,0444	0,8207	0,3311	0,0478
Média	**0,1356**	**0,4529**	**0,0718**	**0,4149**	**0,0321**	**0,9409**	**0,0326**	**0,9386**	**0,1376**	**0,3809**

Tabela 81 - Precisão e Lembrança para expansão vetorial com *k*=25, janela de 50, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0583	0,0566	0,0565	0,0796
0,2	0,0971	0,0648	0,0566	0,0567	0,0986
0,3	0,1237	0,0719	0,0571	0,0569	0,1189
0,4	0,1500	0,0810	0,0574	0,0572	0,1517
0,5	0,1892	0,1014	0,0585	0,0589	0,1541
0,6	0,1977	0,1215	0,0605	0,0600	0,1591
0,7	0,1589	0,0987	0,0702	0,0722	0,1219
0,8	0,1420	0,0659	0,0787	0,0843	0,0836
Média	**0,1421**	**0,0829**	**0,0619**	**0,0628**	**0,1209**

Tabela 82 - Medida *F* para expansão vetorial com *k*=25, janela de 50, e limiar de 0.1 até 0.8

k=30	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0301	0,9245	0,0291	0,9822	0,0290	0,9822	0,0417	0,8749
0,2	0,0517	0,7897	0,0336	0,8104	0,0291	0,9777	0,0291	0,9777	0,0526	0,7333
0,3	0,0687	0,6111	0,0363	0,6445	0,0292	0,9777	0,0291	0,9739	0,0665	0,5618
0,4	0,0887	0,5031	0,0409	0,4551	0,0293	0,9700	0,0292	0,9700	0,0891	0,4328
0,5	0,1238	0,3998	0,0559	0,2839	0,0296	0,9578	0,0298	0,9501	0,1097	0,2466
0,6	0,1746	0,2154	0,0743	0,1615	0,0314	0,9492	0,0314	0,9555	0,1505	0,1518
0,7	0,2302	0,1266	0,1015	0,1009	0,0351	0,9114	0,0358	0,9157	0,2404	0,0875
0,8	0,3376	0,0851	0,1625	0,0437	0,0394	0,8129	0,0398	0,8319	0,3395	0,0501
Média	**0,1395**	**0,4510**	**0,0669**	**0,4281**	**0,0315**	**0,9424**	**0,0317**	**0,9446**	**0,1363**	**0,3924**

Tabela 83 - Precisão e Lembrança para expansão vetorial com *k*=30, janela de 50, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0778	0,0583	0,0564	0,0564	0,0795
0,2	0,0971	0,0645	0,0564	0,0565	0,0982
0,3	0,1234	0,0688	0,0568	0,0566	0,1189
0,4	0,1508	0,0751	0,0568	0,0567	0,1477
0,5	0,1891	0,0934	0,0575	0,0578	0,1519
0,6	0,1929	0,1018	0,0609	0,0608	0,1511
0,7	0,1633	0,1012	0,0677	0,0689	0,1283
0,8	0,1360	0,0689	0,0751	0,0759	0,0873
Média	**0,1413**	**0,0790**	**0,0609**	**0,0612**	**0,1204**

Tabela 84 - Medida *F* para expansão vetorial com *k*=30, janela de 50, e limiar de 0.1 até 0.8

k=35	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0299	0,9227	0,0290	0,9833	0,0290	0,9822	0,0416	0,8728
0,2	0,0516	0,7945	0,0337	0,8089	0,0290	0,9822	0,0291	0,9822	0,0529	0,7359
0,3	0,0678	0,6105	0,0390	0,6422	0,0291	0,9777	0,0291	0,9739	0,0667	0,5631
0,4	0,0881	0,5011	0,0546	0,4180	0,0292	0,9700	0,0291	0,9700	0,0890	0,4338
0,5	0,1222	0,4046	0,0711	0,2372	0,0295	0,9578	0,0297	0,9539	0,1126	0,2549
0,6	0,1654	0,2648	0,1054	0,1720	0,0311	0,9503	0,0314	0,9684	0,1643	0,1682
0,7	0,2297	0,1300	0,1030	0,0909	0,0338	0,9286	0,0342	0,9181	0,2656	0,0947
0,8	0,3019	0,0802	0,1313	0,0405	0,0366	0,8276	0,0385	0,8847	0,3365	0,0478
Média	**0,1334**	**0,4578**	**0,0710**	**0,4165**	**0,0309**	**0,9472**	**0,0313**	**0,9542**	**0,1411**	**0,3964**

Tabela 85 - Precisão e Lembrança para expansão vetorial com *k*=35, janela de 50, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0777	0,0580	0,0563	0,0563	0,0794
0,2	0,0969	0,0647	0,0564	0,0565	0,0988
0,3	0,1220	0,0735	0,0566	0,0564	0,1193
0,4	0,1498	0,0966	0,0566	0,0565	0,1477
0,5	0,1877	0,1094	0,0572	0,0575	0,1562
0,6	0,2036	0,1307	0,0603	0,0607	0,1662
0,7	0,1660	0,0966	0,0651	0,0660	0,1397
0,8	0,1267	0,0619	0,0701	0,0739	0,0838
Média	**0,1413**	**0,0864**	**0,0598**	**0,0605**	**0,1239**

Tabela 86 - Medida *F* para expansão vetorial com *k*=35, janela de 50, e limiar de 0.1 até 0.8

k=40	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8782	0,0298	0,9308	0,0289	0,9833	0,0288	0,9833	0,0414	0,8773
0,2	0,0518	0,7972	0,0334	0,8361	0,0289	0,9822	0,0289	0,9822	0,0516	0,7371
0,3	0,0685	0,6060	0,0374	0,6649	0,0290	0,9777	0,0290	0,9784	0,0648	0,5852
0,4	0,0875	0,4984	0,0575	0,4404	0,0291	0,9739	0,0291	0,9784	0,0855	0,4489
0,5	0,1239	0,4011	0,0688	0,2455	0,0305	0,9771	0,0307	0,9900	0,1192	0,3028
0,6	0,1675	0,2618	0,0952	0,1651	0,0312	0,9690	0,0310	0,9854	0,1537	0,1892
0,7	0,2260	0,1309	0,1160	0,0978	0,0334	0,9273	0,0332	0,9413	0,2229	0,1063
0,8	0,3002	0,0771	0,1698	0,0480	0,0356	0,8214	0,0357	0,8636	0,3150	0,0571
Média	**0,1333**	**0,4563**	**0,0760**	**0,4286**	**0,0308**	**0,9515**	**0,0308**	**0,9628**	**0,1318**	**0,4130**

Tabela 87 - Precisão e Lembrança para expansão vetorial com *k*=40, janela de 50, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0777	0,0577	0,0561	0,0560	0,0790
0,2	0,0972	0,0643	0,0562	0,0562	0,0965
0,3	0,1231	0,0709	0,0564	0,0562	0,1166
0,4	0,1489	0,1017	0,0565	0,0565	0,1437
0,5	0,1893	0,1075	0,0592	0,0595	0,1711
0,6	0,2043	0,1208	0,0604	0,0601	0,1696
0,7	0,1658	0,1062	0,0644	0,0641	0,1439
0,8	0,1227	0,0749	0,0683	0,0686	0,0967
Média	**0,1411**	**0,0880**	**0,0597**	**0,0597**	**0,1272**

Tabela 88 - Medida *F* para expansão vetorial com *k*=40, janela de 50, e limiar de 0.1 até 0.8

k=45	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8782	0,0296	0,9313	0,0288	0,9833	0,0288	0,9833	0,0413	0,8790
0,2	0,0515	0,7911	0,0329	0,8267	0,0289	0,9822	0,0288	0,9822	0,0512	0,7558
0,3	0,0680	0,6060	0,0375	0,6631	0,0290	0,9784	0,0289	0,9784	0,0635	0,5739
0,4	0,0875	0,5024	0,0560	0,4334	0,0291	0,9784	0,0290	0,9784	0,0845	0,4564
0,5	0,1234	0,4066	0,0693	0,2392	0,0308	0,9900	0,0306	0,9900	0,1090	0,2935
0,6	0,1604	0,2608	0,0874	0,1532	0,0312	0,9823	0,0309	0,9871	0,1532	0,2029
0,7	0,2239	0,1313	0,1266	0,1003	0,0325	0,9395	0,0329	0,9533	0,2075	0,1137
0,8	0,2979	0,0786	0,1328	0,0295	0,0357	0,8574	0,0332	0,8596	0,3161	0,0615
Média	**0,1316**	**0,4569**	**0,0715**	**0,4221**	**0,0307**	**0,9614**	**0,0304**	**0,9641**	**0,1283**	**0,4171**

Tabela 89 - Precisão e Lembrança para expansão vetorial com *k*=45, janela de 50, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0574	0,0560	0,0560	0,0790
0,2	0,0966	0,0633	0,0561	0,0560	0,0960
0,3	0,1223	0,0710	0,0563	0,0561	0,1143
0,4	0,1490	0,0992	0,0564	0,0564	0,1425
0,5	0,1893	0,1075	0,0597	0,0594	0,1590
0,6	0,1986	0,1113	0,0605	0,0598	0,1746
0,7	0,1655	0,1119	0,0629	0,0637	0,1469
0,8	0,1244	0,0483	0,0685	0,0638	0,1030
Média	**0,1404**	**0,0837**	**0,0595**	**0,0589**	**0,1269**

Tabela 90 - Medida *F* para expansão vetorial com *k*=45, janela de 50, e limiar de 0.1 até 0.8

k=50	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8827	0,0294	0,9324	0,0288	0,9833	0,0287	0,9833	0,0412	0,8790
0,2	0,0517	0,7986	0,0327	0,8361	0,0288	0,9833	0,0288	0,9833	0,0514	0,7556
0,3	0,0682	0,6141	0,0368	0,6706	0,0289	0,9795	0,0288	0,9795	0,0638	0,5792
0,4	0,0869	0,5058	0,0515	0,4400	0,0290	0,9795	0,0289	0,9795	0,0843	0,4481
0,5	0,1212	0,4090	0,0629	0,2356	0,0305	0,9960	0,0305	0,9912	0,1110	0,2935
0,6	0,1618	0,2722	0,0868	0,1547	0,0308	0,9841	0,0309	0,9841	0,1529	0,2002
0,7	0,2289	0,1399	0,0977	0,0917	0,0311	0,9444	0,0318	0,9455	0,2213	0,1273
0,8	0,3005	0,0840	0,1189	0,0314	0,0337	0,8719	0,0330	0,8718	0,3252	0,0648
Média	**0,1325**	**0,4633**	**0,0646**	**0,4241**	**0,0302**	**0,9652**	**0,0302**	**0,9648**	**0,1314**	**0,4185**

Tabela 91 - Precisão e Lembrança para expansão vetorial com *k*=50, janela de 50, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0570	0,0559	0,0559	0,0788
0,2	0,0971	0,0629	0,0560	0,0560	0,0963
0,3	0,1228	0,0697	0,0562	0,0560	0,1149
0,4	0,1483	0,0923	0,0563	0,0562	0,1419
0,5	0,1870	0,0993	0,0592	0,0592	0,1611
0,6	0,2030	0,1112	0,0597	0,0599	0,1734
0,7	0,1736	0,0946	0,0602	0,0615	0,1616
0,8	0,1313	0,0496	0,0648	0,0636	0,1081
Média	**0,1426**	**0,0796**	**0,0585**	**0,0585**	**0,1295**

Tabela 92 - Medida *F* para expansão vetorial com *k*=50, janela de 50, e limiar de 0.1 até 0.8

Janela de 100

k=5	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8735	0,0318	0,8652	0,0310	0,9466	0,0309	0,9404	0,0416	0,8659
0,2	0,0538	0,7767	0,0348	0,7779	0,0310	0,9092	0,0310	0,9106	0,0516	0,7571
0,3	0,0705	0,6064	0,0395	0,6917	0,0317	0,8879	0,0317	0,8855	0,0634	0,5587
0,4	0,0931	0,4613	0,0454	0,5902	0,0327	0,8756	0,0319	0,8688	0,0845	0,3932
0,5	0,1287	0,3228	0,0584	0,4623	0,0346	0,8634	0,0345	0,8643	0,1054	0,2393
0,6	0,1597	0,1734	0,0777	0,3386	0,0379	0,8106	0,0385	0,8424	0,1325	0,1530
0,7	0,2375	0,1105	0,1239	0,2072	0,0491	0,7733	0,0511	0,7347	0,1704	0,1066
0,8	0,3361	0,0764	0,1403	0,0798	0,0670	0,4727	0,0711	0,4989	0,3006	0,0668
Média	**0,1400**	**0,4251**	**0,0690**	**0,5016**	**0,0394**	**0,8174**	**0,0401**	**0,8182**	**0,1187**	**0,3926**

Tabela 93 - Precisão e Lembrança para expansão vetorial com *k*=5, janela de 100, e limiar de 0.1 até 0.8

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0614	0,0600	0,0598	0,0793
0,2	0,1006	0,0667	0,0599	0,0599	0,0966
0,3	0,1264	0,0747	0,0612	0,0612	0,1139
0,4	0,1549	0,0844	0,0630	0,0616	0,1391
0,5	0,1840	0,1037	0,0665	0,0663	0,1463
0,6	0,1663	0,1264	0,0723	0,0737	0,1420
0,7	0,1508	0,1551	0,0924	0,0955	0,1312
0,8	0,1245	0,1017	0,1174	0,1245	0,1093
Média	**0,1357**	**0,0968**	**0,0741**	**0,0753**	**0,1197**

Tabela 94 - Medida *F* para expansão vetorial com *k*=5, janela de 100, e limiar de 0.1 até 0.8

k=10	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8755	0,0311	0,8934	0,0304	0,9592	0,0303	0,9581	0,0417	0,8674
0,2	0,0528	0,7890	0,0329	0,7968	0,0306	0,9393	0,0305	0,9445	0,0520	0,7472
0,3	0,0697	0,6209	0,0344	0,6819	0,0310	0,9187	0,0308	0,9238	0,0645	0,5606
0,4	0,0897	0,4756	0,0372	0,6050	0,0312	0,8973	0,0309	0,9001	0,0859	0,3852
0,5	0,1233	0,3408	0,0429	0,5123	0,0324	0,8524	0,0317	0,8517	0,1019	0,2185
0,6	0,1593	0,1824	0,0510	0,4411	0,0337	0,8179	0,0340	0,8240	0,1432	0,1551
0,7	0,2291	0,1138	0,0789	0,3341	0,0441	0,7987	0,0410	0,8105	0,2080	0,0998
0,8	0,3769	0,0669	0,1295	0,1658	0,0512	0,6664	0,0532	0,7184	0,3166	0,0552
Média	**0,1427**	**0,4331**	**0,0547**	**0,5538**	**0,0356**	**0,8562**	**0,0353**	**0,8664**	**0,1267**	**0,3861**

Tabela 95 - Precisão e Lembrança para expansão vetorial com *k*=10, janela de 100, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0600	0,0590	0,0587	0,0795
0,2	0,0990	0,0632	0,0593	0,0591	0,0972
0,3	0,1253	0,0656	0,0599	0,0596	0,1157
0,4	0,1509	0,0701	0,0603	0,0597	0,1404
0,5	0,1811	0,0791	0,0625	0,0610	0,1390
0,6	0,1700	0,0915	0,0647	0,0653	0,1489
0,7	0,1521	0,1276	0,0836	0,0780	0,1349
0,8	0,1136	0,1454	0,0951	0,0990	0,0940
Média	**0,1338**	**0,0878**	**0,0680**	**0,0675**	**0,1187**

Tabela 96 - Medida *F* para expansão vetorial com *k*=10, janela de 100, e limiar de 0.1 até 0.8

k=15	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8767	0,0307	0,9081	0,0297	0,9762	0,0296	0,9654	0,0417	0,8692
0,2	0,0532	0,7898	0,0331	0,8164	0,0297	0,9671	0,0297	0,9518	0,0519	0,7419
0,3	0,0694	0,6017	0,0358	0,7155	0,0301	0,9594	0,0299	0,9350	0,0642	0,5669
0,4	0,0898	0,4750	0,0414	0,6358	0,0305	0,9518	0,0302	0,9228	0,0845	0,3796
0,5	0,1230	0,3464	0,0467	0,5316	0,0310	0,9256	0,0309	0,8973	0,1093	0,2443
0,6	0,1608	0,1916	0,0429	0,4261	0,0328	0,8841	0,0318	0,8579	0,1415	0,1586
0,7	0,2161	0,1155	0,0562	0,3240	0,0410	0,8649	0,0383	0,8482	0,1900	0,0850
0,8	0,3502	0,0792	0,0837	0,1915	0,0488	0,7532	0,0483	0,7864	0,3266	0,0575
Média	**0,1379**	**0,4345**	**0,0463**	**0,5686**	**0,0342**	**0,9103**	**0,0336**	**0,8956**	**0,1262**	**0,3879**

Tabela 97 - Precisão e Lembrança para expansão vetorial com *k*=15, janela de 100, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0782	0,0593	0,0576	0,0575	0,0795
0,2	0,0996	0,0636	0,0576	0,0576	0,0971
0,3	0,1245	0,0682	0,0583	0,0579	0,1153
0,4	0,1511	0,0777	0,0591	0,0585	0,1382
0,5	0,1816	0,0859	0,0600	0,0597	0,1511
0,6	0,1748	0,0779	0,0632	0,0613	0,1496
0,7	0,1505	0,0958	0,0783	0,0734	0,1175
0,8	0,1292	0,1165	0,0916	0,0910	0,0978
Média	**0,1362**	**0,0806**	**0,0657**	**0,0646**	**0,1182**

Tabela 98 - Medida *F* para expansão vetorial com *k*=15, janela de 100, e limiar de 0.1 até 0.8

k=20	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8767	0,0302	0,9220	0,0293	0,9773	0,0292	0,9773	0,0418	0,8739
0,2	0,0524	0,7903	0,0328	0,8426	0,0293	0,9682	0,0293	0,9682	0,0519	0,7335
0,3	0,0693	0,6044	0,0344	0,7324	0,0297	0,9682	0,0294	0,9644	0,0650	0,5647
0,4	0,0888	0,4852	0,0379	0,6295	0,0299	0,9644	0,0297	0,9567	0,0860	0,3962
0,5	0,1252	0,3653	0,0437	0,5309	0,0303	0,9483	0,0300	0,9444	0,1082	0,2558
0,6	0,1769	0,2042	0,0403	0,4330	0,0315	0,9081	0,0315	0,9247	0,1455	0,1594
0,7	0,2181	0,1211	0,0478	0,3151	0,0370	0,9068	0,0353	0,9108	0,2006	0,0936
0,8	0,3037	0,0729	0,0694	0,1950	0,0447	0,8318	0,0421	0,8015	0,3073	0,0537
Média	**0,1344**	**0,4400**	**0,0421**	**0,5751**	**0,0327**	**0,9341**	**0,0321**	**0,9310**	**0,1258**	**0,3914**

Tabela 99 - Precisão e Lembrança para expansão vetorial com *k*=20, janela de 100, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0586	0,0568	0,0568	0,0798
0,2	0,0982	0,0632	0,0568	0,0569	0,0969
0,3	0,1244	0,0658	0,0576	0,0571	0,1166
0,4	0,1502	0,0715	0,0580	0,0576	0,1413
0,5	0,1864	0,0808	0,0588	0,0581	0,1521
0,6	0,1896	0,0738	0,0608	0,0610	0,1521
0,7	0,1557	0,0830	0,0711	0,0681	0,1277
0,8	0,1175	0,1024	0,0848	0,0800	0,0915
Média	**0,1375**	**0,0749**	**0,0631**	**0,0619**	**0,1198**

Tabela 100 - Medida *F* para expansão vetorial com *k*=20, janela de 100, e limiar de 0.1 até 0.8

k=25	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8771	0,0298	0,9242	0,0292	0,9784	0,0291	0,9784	0,0418	0,8749
0,2	0,0518	0,7894	0,0323	0,8536	0,0292	0,9739	0,0292	0,9739	0,0527	0,7418
0,3	0,0697	0,6262	0,0332	0,7414	0,0294	0,9739	0,0293	0,9739	0,0657	0,5592
0,4	0,0880	0,4930	0,0366	0,6529	0,0296	0,9700	0,0295	0,9700	0,0906	0,3884
0,5	0,1273	0,3937	0,0383	0,5511	0,0299	0,9700	0,0304	0,9578	0,1080	0,2461
0,6	0,1708	0,2068	0,0399	0,4597	0,0311	0,9411	0,0312	0,9195	0,1424	0,1608
0,7	0,2133	0,1258	0,0502	0,3445	0,0346	0,9105	0,0377	0,9173	0,2172	0,0980
0,8	0,3276	0,0758	0,0667	0,1913	0,0403	0,8118	0,0435	0,8195	0,2777	0,0537
Média	**0,1362**	**0,4485**	**0,0409**	**0,5898**	**0,0317**	**0,9412**	**0,0325**	**0,9388**	**0,1245**	**0,3904**

Tabela 101 - Precisão e Lembrança para expansão vetorial com *k*=25, janela de 100, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0578	0,0567	0,0565	0,0797
0,2	0,0973	0,0622	0,0567	0,0567	0,0985
0,3	0,1255	0,0635	0,0571	0,0569	0,1176
0,4	0,1493	0,0694	0,0575	0,0572	0,1469
0,5	0,1923	0,0716	0,0580	0,0589	0,1501
0,6	0,1871	0,0734	0,0602	0,0604	0,1511
0,7	0,1583	0,0876	0,0667	0,0724	0,1351
0,8	0,1231	0,0989	0,0768	0,0826	0,0900
Média	**0,1389**	**0,0731**	**0,0612**	**0,0627**	**0,1211**

Tabela 102 - Medida *F* para expansão vetorial com *k*=25, janela de 100, e limiar de 0.1 até 0.8

k=30	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0298	0,9379	0,0291	0,9822	0,0290	0,9822	0,0417	0,8768
0,2	0,0516	0,7904	0,0324	0,8842	0,0291	0,9777	0,0291	0,9777	0,0532	0,7420
0,3	0,0698	0,6265	0,0328	0,7600	0,0292	0,9777	0,0292	0,9739	0,0657	0,5599
0,4	0,0883	0,4956	0,0346	0,6506	0,0293	0,9700	0,0292	0,9700	0,0907	0,3973
0,5	0,1277	0,3987	0,0379	0,5584	0,0296	0,9655	0,0298	0,9501	0,1102	0,2556
0,6	0,1800	0,2131	0,0520	0,4332	0,0313	0,9503	0,0314	0,9544	0,1533	0,1801
0,7	0,2142	0,1269	0,0460	0,3323	0,0349	0,9157	0,0361	0,9173	0,2233	0,1060
0,8	0,3284	0,0787	0,0612	0,1794	0,0380	0,8139	0,0412	0,8771	0,2804	0,0550
Média	**0,1376**	**0,4509**	**0,0408**	**0,5920**	**0,0313**	**0,9441**	**0,0319**	**0,9503**	**0,1273**	**0,3966**

Tabela 103 - Precisão e Lembrança para expansão vetorial com *k*=30, janela de 100, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0779	0,0577	0,0564	0,0564	0,0796
0,2	0,0969	0,0625	0,0564	0,0565	0,0993
0,3	0,1256	0,0628	0,0567	0,0566	0,1175
0,4	0,1498	0,0656	0,0568	0,0567	0,1476
0,5	0,1934	0,0709	0,0575	0,0579	0,1540
0,6	0,1952	0,0928	0,0606	0,0608	0,1656
0,7	0,1594	0,0808	0,0673	0,0694	0,1438
0,8	0,1270	0,0913	0,0726	0,0787	0,0919
Média	**0,1406**	**0,0731**	**0,0606**	**0,0616**	**0,1249**

Tabela 104 - Medida *F* para expansão vetorial com *k*=30, janela de 100, e limiar de 0.1 até 0.8

k=35	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0297	0,9411	0,0290	0,9833	0,0289	0,9822	0,0415	0,8769
0,2	0,0513	0,7881	0,0321	0,8847	0,0290	0,9822	0,0290	0,9822	0,0527	0,7412
0,3	0,0690	0,6259	0,0328	0,7815	0,0291	0,9777	0,0290	0,9739	0,0665	0,5666
0,4	0,0881	0,4999	0,0343	0,6677	0,0291	0,9700	0,0291	0,9700	0,0904	0,3902
0,5	0,1252	0,4016	0,0362	0,5635	0,0295	0,9578	0,0308	0,9731	0,1097	0,2551
0,6	0,1787	0,2235	0,0412	0,4329	0,0312	0,9509	0,0313	0,9672	0,1470	0,1619
0,7	0,2190	0,1249	0,0481	0,2884	0,0343	0,9160	0,0342	0,9226	0,2205	0,0879
0,8	0,3066	0,0801	0,0562	0,1799	0,0389	0,8602	0,0371	0,8409	0,3773	0,0509
Média	**0,1348**	**0,4526**	**0,0388**	**0,5925**	**0,0313**	**0,9498**	**0,0312**	**0,9515**	**0,1382**	**0,3913**

Tabela 105 - Precisão e Lembrança para expansão vetorial com *k*=35, janela de 100, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0777	0,0576	0,0563	0,0562	0,0793
0,2	0,0963	0,0619	0,0564	0,0563	0,0983
0,3	0,1243	0,0630	0,0566	0,0563	0,1190
0,4	0,1497	0,0652	0,0566	0,0565	0,1468
0,5	0,1908	0,0680	0,0572	0,0597	0,1534
0,6	0,1986	0,0752	0,0605	0,0606	0,1541
0,7	0,1591	0,0824	0,0661	0,0660	0,1257
0,8	0,1270	0,0857	0,0745	0,0710	0,0896
Média	**0,1405**	**0,0699**	**0,0605**	**0,0603**	**0,1208**

Tabela 106 - Medida *F* para expansão vetorial com *k*=35, janela de 100, e limiar de 0.1 até 0.8

k=40	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8771	0,0297	0,9474	0,0289	0,9833	0,0288	0,9833	0,0415	0,8775
0,2	0,0515	0,7975	0,0321	0,8992	0,0289	0,9822	0,0289	0,9822	0,0520	0,7386
0,3	0,0683	0,6221	0,0326	0,7990	0,0291	0,9822	0,0290	0,9784	0,0654	0,5720
0,4	0,0871	0,5021	0,0332	0,6852	0,0291	0,9784	0,0291	0,9784	0,0896	0,4067
0,5	0,1228	0,4027	0,0348	0,5745	0,0293	0,9495	0,0306	0,9779	0,1217	0,3021
0,6	0,1690	0,2692	0,0375	0,4394	0,0313	0,9790	0,0309	0,9733	0,1522	0,1789
0,7	0,2287	0,1317	0,0417	0,2773	0,0336	0,9298	0,0333	0,9291	0,2039	0,1067
0,8	0,3071	0,0776	0,0512	0,1554	0,0360	0,8259	0,0353	0,8503	0,3330	0,0589
Média	**0,1344**	**0,4600**	**0,0366**	**0,5972**	**0,0308**	**0,9513**	**0,0307**	**0,9566**	**0,1324**	**0,4052**

Tabela 107 - Precisão e Lembrança para expansão vetorial com *k*=40, janela de 100, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0575	0,0561	0,0560	0,0792
0,2	0,0968	0,0619	0,0562	0,0561	0,0972
0,3	0,1231	0,0626	0,0564	0,0563	0,1174
0,4	0,1484	0,0633	0,0565	0,0566	0,1468
0,5	0,1882	0,0656	0,0568	0,0593	0,1735
0,6	0,2076	0,0692	0,0607	0,0600	0,1645
0,7	0,1672	0,0725	0,0649	0,0642	0,1401
0,8	0,1239	0,0770	0,0690	0,0678	0,1002
Média	**0,1416**	**0,0662**	**0,0596**	**0,0595**	**0,1274**

Tabela 108 - Medida *F* para expansão vetorial com *k*=40, janela de 100, e limiar de 0.1 até 0.8

k=45	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8782	0,0295	0,9530	0,0288	0,9833	0,0288	0,9833	0,0414	0,8779
0,2	0,0516	0,7989	0,0319	0,9004	0,0289	0,9822	0,0288	0,9822	0,0522	0,7519
0,3	0,0683	0,6202	0,0318	0,8114	0,0290	0,9784	0,0289	0,9784	0,0659	0,5840
0,4	0,0876	0,5025	0,0327	0,6885	0,0291	0,9784	0,0291	0,9784	0,0856	0,4423
0,5	0,1228	0,4029	0,0332	0,5564	0,0296	0,9739	0,0306	0,9900	0,1121	0,2823
0,6	0,1667	0,2684	0,0370	0,4368	0,0310	0,9877	0,0308	0,9871	0,1555	0,1825
0,7	0,2231	0,1312	0,0433	0,3144	0,0326	0,9435	0,0328	0,9510	0,2340	0,1138
0,8	0,3027	0,0801	0,0519	0,1692	0,0351	0,8632	0,0351	0,8758	0,3221	0,0622
Média	**0,1329**	**0,4603**	**0,0364**	**0,6038**	**0,0305**	**0,9613**	**0,0306**	**0,9658**	**0,1336**	**0,4121**

Tabela 109 - Precisão e Lembrança para expansão vetorial com *k*=45, janela de 100, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0573	0,0560	0,0559	0,0791
0,2	0,0970	0,0616	0,0561	0,0560	0,0977
0,3	0,1230	0,0613	0,0563	0,0562	0,1185
0,4	0,1492	0,0624	0,0564	0,0564	0,1434
0,5	0,1882	0,0627	0,0574	0,0594	0,1604
0,6	0,2057	0,0682	0,0601	0,0598	0,1680
0,7	0,1652	0,0761	0,0629	0,0634	0,1531
0,8	0,1266	0,0794	0,0674	0,0675	0,1043
Média	**0,1416**	**0,0661**	**0,0591**	**0,0593**	**0,1281**

Tabela 110 - Medida *F* para expansão vetorial com *k*=45, janela de 100, e limiar de 0.1 até 0.8

k=50	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0405	0,8782	0,0293	0,9529	0,0288	0,9833	0,0287	0,9833	0,0413	0,8773
0,2	0,0515	0,7926	0,0316	0,9001	0,0288	0,9833	0,0288	0,9833	0,0520	0,7447
0,3	0,0679	0,6213	0,0312	0,7907	0,0289	0,9795	0,0289	0,9795	0,0656	0,5919
0,4	0,0874	0,5037	0,0323	0,6772	0,0290	0,9795	0,0290	0,9795	0,0829	0,4355
0,5	0,1149	0,3858	0,0332	0,5526	0,0305	0,9960	0,0305	0,9960	0,1186	0,2898
0,6	0,1565	0,2575	0,0384	0,4431	0,0307	0,9889	0,0305	0,9937	0,1522	0,1770
0,7	0,2230	0,1339	0,0445	0,3119	0,0313	0,9528	0,0319	0,9619	0,2308	0,1056
0,8	0,3098	0,0889	0,0525	0,1715	0,0328	0,8696	0,0335	0,8859	0,3742	0,0755
Média	**0,1314**	**0,4577**	**0,0366**	**0,6000**	**0,0301**	**0,9666**	**0,0302**	**0,9704**	**0,1397**	**0,4122**

Tabela 111 - Precisão e Lembrança para expansão vetorial com *k*=50, janela de 100, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0774	0,0569	0,0559	0,0559	0,0789
0,2	0,0967	0,0610	0,0560	0,0559	0,0972
0,3	0,1225	0,0601	0,0562	0,0562	0,1181
0,4	0,1489	0,0617	0,0562	0,0563	0,1393
0,5	0,1771	0,0627	0,0591	0,0591	0,1683
0,6	0,1947	0,0707	0,0596	0,0592	0,1637
0,7	0,1674	0,0779	0,0606	0,0617	0,1449
0,8	0,1381	0,0803	0,0632	0,0646	0,1256
Média	**0,1403**	**0,0664**	**0,0584**	**0,0586**	**0,1295**

Tabela 112 - Medida *F* para expansão vetorial com *k*=50, janela de 100, e limiar de 0.1 até 0.8

Janela de 200

k=5	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8735	0,0356	0,8992	0,0310	0,9466	0,0309	0,9404	0,0415	0,8659
0,2	0,0538	0,7767	0,0452	0,7769	0,0310	0,9092	0,0310	0,9106	0,0516	0,7571
0,3	0,0704	0,6064	0,0522	0,6510	0,0317	0,8879	0,0317	0,8855	0,0634	0,5587
0,4	0,0932	0,4617	0,0645	0,5323	0,0327	0,8756	0,0319	0,8688	0,0845	0,3939
0,5	0,1285	0,3196	0,0782	0,3895	0,0345	0,8634	0,0345	0,8643	0,1049	0,2393
0,6	0,1601	0,1739	0,1039	0,2534	0,0379	0,8106	0,0385	0,8424	0,1323	0,1530
0,7	0,2368	0,1105	0,1170	0,1320	0,0492	0,7726	0,0511	0,7347	0,1750	0,1074
0,8	0,3319	0,0764	0,2076	0,0743	0,0667	0,4713	0,0711	0,4989	0,3001	0,0668
Média	**0,1394**	**0,4248**	**0,0880**	**0,4636**	**0,0393**	**0,8171**	**0,0401**	**0,8182**	**0,1191**	**0,3928**

Tabela 113 - Precisão e Lembrança para expansão vetorial com *k*=5, janela de 200, e limiar de 0.1 até 0.8

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0686	0,0600	0,0598	0,0793
0,2	0,1005	0,0854	0,0599	0,0599	0,0965
0,3	0,1261	0,0966	0,0612	0,0612	0,1138
0,4	0,1550	0,1150	0,0630	0,0616	0,1391
0,5	0,1833	0,1303	0,0664	0,0663	0,1458
0,6	0,1667	0,1474	0,0725	0,0737	0,1419
0,7	0,1507	0,1240	0,0926	0,0955	0,1331
0,8	0,1242	0,1094	0,1168	0,1245	0,1093
Média	**0,1356**	**0,1096**	**0,0740**	**0,0753**	**0,1199**

Tabela 114 - Medida *F* para expansão vetorial com *k*=5, janela de 200, e limiar de 0.1 até 0.8

k=10	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8755	0,0347	0,9219	0,0304	0,9592	0,0303	0,9581	0,0417	0,8674
0,2	0,0527	0,7852	0,0393	0,8278	0,0306	0,9393	0,0305	0,9445	0,0521	0,7535
0,3	0,0696	0,6209	0,0441	0,7123	0,0309	0,9187	0,0308	0,9238	0,0644	0,5606
0,4	0,0895	0,4750	0,0557	0,6319	0,0312	0,8973	0,0309	0,9001	0,0840	0,3778
0,5	0,1236	0,3404	0,0681	0,5174	0,0324	0,8524	0,0317	0,8517	0,1019	0,2207
0,6	0,1588	0,1824	0,0811	0,3396	0,0334	0,8179	0,0340	0,8240	0,1427	0,1551
0,7	0,2301	0,1138	0,0957	0,1839	0,0434	0,7970	0,0410	0,8105	0,2113	0,1006
0,8	0,3769	0,0669	0,1966	0,0780	0,0517	0,6830	0,0532	0,7151	0,2958	0,0552
Média	**0,1427**	**0,4325**	**0,0769**	**0,5266**	**0,0355**	**0,8581**	**0,0353**	**0,8660**	**0,1242**	**0,3864**

Tabela 115 - Precisão e Lembrança para expansão vetorial com *k*=10, janela de 200, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0781	0,0668	0,0590	0,0587	0,0795
0,2	0,0987	0,0751	0,0593	0,0591	0,0974
0,3	0,1252	0,0830	0,0599	0,0596	0,1156
0,4	0,1506	0,1025	0,0604	0,0597	0,1375
0,5	0,1814	0,1203	0,0624	0,0611	0,1394
0,6	0,1698	0,1310	0,0642	0,0653	0,1486
0,7	0,1523	0,1259	0,0824	0,0781	0,1363
0,8	0,1136	0,1117	0,0961	0,0990	0,0930
Média	**0,1337**	**0,1020**	**0,0680**	**0,0676**	**0,1184**

Tabela 116 - Medida *F* para expansão vetorial com *k*=10, janela de 200, e limiar de 0.1 até 0.8

k=15	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0409	0,8767	0,0313	0,9541	0,0297	0,9762	0,0296	0,9654	0,0417	0,8692
0,2	0,0527	0,7930	0,0361	0,8876	0,0297	0,9671	0,0297	0,9518	0,0518	0,7414
0,3	0,0688	0,6055	0,0396	0,8000	0,0301	0,9594	0,0299	0,9350	0,0643	0,5679
0,4	0,0889	0,4792	0,0453	0,6699	0,0305	0,9518	0,0302	0,9228	0,0845	0,3807
0,5	0,1234	0,3586	0,0519	0,5592	0,0310	0,9256	0,0309	0,8973	0,1093	0,2443
0,6	0,1605	0,1973	0,0680	0,4560	0,0333	0,8829	0,0318	0,8579	0,1406	0,1582
0,7	0,2215	0,1190	0,1010	0,2700	0,0410	0,8691	0,0383	0,8482	0,1931	0,0915
0,8	0,3233	0,0809	0,1744	0,1187	0,0493	0,7638	0,0483	0,7863	0,3272	0,0575
Média	**0,1350**	**0,4388**	**0,0684**	**0,5894**	**0,0343**	**0,9120**	**0,0336**	**0,8956**	**0,1266**	**0,3888**

Tabela 117 - Precisão e Lembrança para expansão vetorial com *k*=15, janela de 200, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0782	0,0607	0,0576	0,0575	0,0795
0,2	0,0988	0,0693	0,0576	0,0576	0,0969
0,3	0,1236	0,0754	0,0584	0,0579	0,1155
0,4	0,1499	0,0849	0,0591	0,0585	0,1383
0,5	0,1836	0,0949	0,0601	0,0597	0,1510
0,6	0,1770	0,1183	0,0641	0,0613	0,1489
0,7	0,1548	0,1470	0,0783	0,0734	0,1242
0,8	0,1294	0,1412	0,0925	0,0910	0,0978
Média	**0,1369**	**0,0990**	**0,0660**	**0,0646**	**0,1190**

Tabela 118 - Medida *F* para expansão vetorial com *k*=15, janela de 200, e limiar de 0.1 até 0.8

k=20	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8767	0,0306	0,9669	0,0293	0,9773	0,0292	0,9773	0,0418	0,8739
0,2	0,0524	0,7916	0,0339	0,9313	0,0293	0,9682	0,0293	0,9682	0,0517	0,7309
0,3	0,0693	0,6047	0,0360	0,8528	0,0297	0,9682	0,0295	0,9644	0,0651	0,5634
0,4	0,0888	0,4852	0,0404	0,7417	0,0299	0,9644	0,0297	0,9567	0,0857	0,3926
0,5	0,1258	0,3663	0,0468	0,6235	0,0303	0,9483	0,0300	0,9444	0,1080	0,2537
0,6	0,1742	0,2042	0,0582	0,5181	0,0315	0,9081	0,0316	0,9258	0,1456	0,1571
0,7	0,2184	0,1211	0,0698	0,3371	0,0369	0,9061	0,0354	0,9108	0,2131	0,0912
0,8	0,2985	0,0729	0,1439	0,1483	0,0436	0,8310	0,0421	0,8015	0,3153	0,0537
Média	**0,1335**	**0,4403**	**0,0575**	**0,6400**	**0,0326**	**0,9339**	**0,0321**	**0,9311**	**0,1283**	**0,3896**

Tabela 119 - Precisão e Lembrança para expansão vetorial com *k*=20, janela de 200, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0593	0,0568	0,0568	0,0798
0,2	0,0982	0,0655	0,0568	0,0569	0,0966
0,3	0,1244	0,0691	0,0576	0,0572	0,1168
0,4	0,1502	0,0766	0,0580	0,0576	0,1407
0,5	0,1872	0,0870	0,0588	0,0581	0,1515
0,6	0,1880	0,1047	0,0608	0,0611	0,1511
0,7	0,1558	0,1157	0,0710	0,0681	0,1278
0,8	0,1171	0,1461	0,0828	0,0800	0,0918
Média	**0,1374**	**0,0905**	**0,0628**	**0,0620**	**0,1195**

Tabela 120 - Medida *F* para expansão vetorial com *k*=20, janela de 200, e limiar de 0.1 até 0.8

k=25	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0408	0,8771	0,0304	0,9655	0,0292	0,9784	0,0291	0,9784	0,0417	0,8749
0,2	0,0519	0,7913	0,0332	0,9398	0,0292	0,9739	0,0292	0,9739	0,0529	0,7430
0,3	0,0698	0,6262	0,0348	0,8907	0,0294	0,9739	0,0293	0,9739	0,0656	0,5544
0,4	0,0879	0,4930	0,0358	0,7806	0,0296	0,9700	0,0295	0,9700	0,0904	0,3844
0,5	0,1271	0,3947	0,0408	0,6819	0,0299	0,9700	0,0304	0,9578	0,1074	0,2433
0,6	0,1754	0,2106	0,0459	0,5579	0,0311	0,9411	0,0312	0,9195	0,1407	0,1541
0,7	0,2144	0,1258	0,0639	0,4305	0,0345	0,9101	0,0377	0,9173	0,2115	0,0980
0,8	0,3212	0,0740	0,0770	0,2081	0,0401	0,8141	0,0438	0,8271	0,2811	0,0537
Média	**0,1361**	**0,4491**	**0,0452**	**0,6819**	**0,0316**	**0,9414**	**0,0325**	**0,9397**	**0,1239**	**0,3882**

Tabela 121 - Precisão e Lembrança para expansão vetorial com *k*=25, janela de 200, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0780	0,0589	0,0567	0,0671	0,0797
0,2	0,0974	0,0641	0,0566	0,0800	0,0988
0,3	0,1256	0,0669	0,0571	0,0954	0,1173
0,4	0,1493	0,0685	0,0574	0,1137	0,1464
0,5	0,1923	0,0769	0,0580	0,1378	0,1491
0,6	0,1914	0,0848	0,0602	0,1655	0,1470
0,7	0,1586	0,1112	0,0664	0,2270	0,1339
0,8	0,1203	0,1124	0,0765	0,2977	0,0902
Média	**0,1391**	**0,0805**	**0,0611**	**0,1480**	**0,1203**

Tabela 122 - Medida *F* para expansão vetorial com *k*=25, janela de 200, e limiar de 0.1 até 0.8

k=30	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0302	0,9662	0,0291	0,9822	0,0290	0,9822	0,0417	0,8774
0,2	0,0516	0,7904	0,0327	0,9388	0,0291	0,9777	0,0291	0,9777	0,0529	0,7406
0,3	0,0699	0,6265	0,0344	0,9047	0,0292	0,9777	0,0291	0,9739	0,0655	0,5556
0,4	0,0881	0,4956	0,0366	0,8437	0,0293	0,9700	0,0292	0,9700	0,0897	0,3960
0,5	0,1274	0,3987	0,0406	0,7581	0,0296	0,9655	0,0298	0,9501	0,1097	0,2556
0,6	0,1813	0,2139	0,0462	0,6373	0,0313	0,9503	0,0313	0,9544	0,1545	0,1732
0,7	0,2107	0,1242	0,0587	0,4573	0,0349	0,9157	0,0361	0,9184	0,2234	0,1037
0,8	0,3209	0,0771	0,0781	0,2682	0,0381	0,8139	0,0417	0,8760	0,2892	0,0514
Média	**0,1363**	**0,4504**	**0,0447**	**0,7218**	**0,0313**	**0,9441**	**0,0319**	**0,9503**	**0,1283**	**0,3942**

Tabela 123 - Precisão e Lembrança para expansão vetorial com *k*=30, janela de 200, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0779	0,0586	0,0564	0,0564	0,0797
0,2	0,0969	0,0633	0,0564	0,0565	0,0987
0,3	0,1257	0,0663	0,0567	0,0566	0,1172
0,4	0,1496	0,0701	0,0568	0,0567	0,1463
0,5	0,1931	0,0771	0,0575	0,0577	0,1535
0,6	0,1962	0,0861	0,0606	0,0607	0,1633
0,7	0,1563	0,1040	0,0672	0,0694	0,1416
0,8	0,1243	0,1209	0,0728	0,0797	0,0873
Média	**0,1400**	**0,0808**	**0,0606**	**0,0617**	**0,1235**

Tabela 124 - Medida *F* para expansão vetorial com *k*=30, janela de 200, e limiar de 0.1 até 0.8

k=35	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0407	0,8771	0,0301	0,9657	0,0290	0,9833	0,0289	0,9822	0,0415	0,8769
0,2	0,0514	0,7943	0,0327	0,9412	0,0291	0,9822	0,0290	0,9822	0,0526	0,7406
0,3	0,0692	0,6321	0,0337	0,9064	0,0291	0,9777	0,0290	0,9739	0,0665	0,5631
0,4	0,0881	0,4987	0,0358	0,8578	0,0291	0,9700	0,0291	0,9700	0,0904	0,3904
0,5	0,1259	0,4002	0,0397	0,7920	0,0295	0,9578	0,0307	0,9690	0,1099	0,2545
0,6	0,1795	0,2202	0,0462	0,6835	0,0312	0,9509	0,0313	0,9672	0,1515	0,1667
0,7	0,2170	0,1242	0,0570	0,4806	0,0343	0,9160	0,0345	0,9226	0,2125	0,0866
0,8	0,3113	0,0801	0,0781	0,2751	0,0383	0,8584	0,0389	0,8884	0,3773	0,0517
Média	**0,1354**	**0,4534**	**0,0442**	**0,7378**	**0,0312**	**0,9496**	**0,0314**	**0,9570**	**0,1378**	**0,3913**

Tabela 125 - Precisão e Lembrança para expansão vetorial com *k*=35, janela de 200, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0777	0,0585	0,0563	0,0562	0,0793
0,2	0,0965	0,0632	0,0564	0,0563	0,0982
0,3	0,1247	0,0650	0,0566	0,0563	0,1189
0,4	0,1497	0,0686	0,0566	0,0565	0,1468
0,5	0,1915	0,0755	0,0572	0,0596	0,1535
0,6	0,1978	0,0865	0,0604	0,0607	0,1587
0,7	0,1580	0,1019	0,0661	0,0665	0,1231
0,8	0,1274	0,1216	0,0733	0,0745	0,0909
Média	**0,1404**	**0,0801**	**0,0604**	**0,0608**	**0,1212**

Tabela 126 - Medida *F* para expansão vetorial com *k*=35, janela de 200, e limiar de 0.1 até 0.8

k=40	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8771	0,0300	0,9676	0,0289	0,9833	0,0288	0,9833	0,0414	0,8775
0,2	0,0517	0,8001	0,0325	0,9421	0,0289	0,9822	0,0289	0,9822	0,0517	0,7403
0,3	0,0688	0,6296	0,0334	0,9129	0,0291	0,9822	0,0290	0,9784	0,0645	0,5665
0,4	0,0876	0,5025	0,0358	0,8665	0,0291	0,9784	0,0291	0,9784	0,0900	0,4103
0,5	0,1227	0,4023	0,0397	0,7873	0,0293	0,9495	0,0306	0,9779	0,1169	0,2918
0,6	0,1693	0,2651	0,0443	0,7070	0,0313	0,9790	0,0309	0,9733	0,1452	0,1742
0,7	0,2186	0,1302	0,0537	0,5687	0,0337	0,9298	0,0333	0,9291	0,1874	0,1032
0,8	0,2977	0,0776	0,0752	0,3135	0,0361	0,8259	0,0348	0,8455	0,3353	0,0580
Média	**0,1321**	**0,4606**	**0,0431**	**0,7582**	**0,0308**	**0,9513**	**0,0307**	**0,9560**	**0,1290**	**0,4027**

Tabela 127 - Precisão e Lembrança para expansão vetorial com *k*=40, janela de 200, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0582	0,0561	0,0560	0,0791
0,2	0,0971	0,0628	0,0562	0,0561	0,0966
0,3	0,1240	0,0645	0,0564	0,0563	0,1158
0,4	0,1492	0,0688	0,0565	0,0566	0,1477
0,5	0,1880	0,0755	0,0568	0,0593	0,1669
0,6	0,2067	0,0833	0,0607	0,0600	0,1584
0,7	0,1632	0,0981	0,0650	0,0642	0,1331
0,8	0,1232	0,1214	0,0691	0,0669	0,0989
Média	**0,1411**	**0,0791**	**0,0596**	**0,0594**	**0,1245**

Tabela 128 - Medida *F* para expansão vetorial com *k*=40, janela de 200, e limiar de 0.1 até 0.8

k=45	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0406	0,8782	0,0299	0,9676	0,0288	0,9833	0,0288	0,9833	0,0414	0,8786
0,2	0,0515	0,8057	0,0322	0,9450	0,0289	0,9822	0,0288	0,9822	0,0516	0,7454
0,3	0,0684	0,6276	0,0332	0,9219	0,0290	0,9784	0,0289	0,9784	0,0644	0,5706
0,4	0,0862	0,5014	0,0353	0,8907	0,0290	0,9784	0,0291	0,9784	0,0855	0,4447
0,5	0,1224	0,4037	0,0384	0,8116	0,0296	0,9739	0,0306	0,9900	0,1184	0,2992
0,6	0,1673	0,2654	0,0448	0,7405	0,0310	0,9877	0,0308	0,9871	0,1433	0,1707
0,7	0,2290	0,1317	0,0527	0,6137	0,0325	0,9435	0,0325	0,9462	0,2016	0,1008
0,8	0,3062	0,0884	0,0834	0,3516	0,0350	0,8636	0,0343	0,8710	0,3539	0,0678
Média	**0,1340**	**0,4627**	**0,0437**	**0,7803**	**0,0305**	**0,9614**	**0,0305**	**0,9646**	**0,1325**	**0,4097**

Tabela 129 - Precisão e Lembrança para expansão vetorial com *k*=45, janela de 200, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0776	0,0579	0,0560	0,0559	0,0790
0,2	0,0969	0,0622	0,0561	0,0560	0,0965
0,3	0,1234	0,0642	0,0563	0,0562	0,1158
0,4	0,1471	0,0680	0,0564	0,0564	0,1435
0,5	0,1878	0,0734	0,0574	0,0594	0,1697
0,6	0,2052	0,0844	0,0601	0,0598	0,1558
0,7	0,1672	0,0971	0,0629	0,0629	0,1344
0,8	0,1372	0,1348	0,0673	0,0660	0,1137
Média	**0,1428**	**0,0803**	**0,0591**	**0,0591**	**0,1260**

Tabela 130 - Medida *F* para expansão vetorial com *k*=45, janela de 200, e limiar de 0.1 até 0.8

k=50	LRD		Phi-squared		MI		VMI		Z score	
Limiar	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
0,1	0,0405	0,8782	0,0298	0,9699	0,0288	0,9833	0,0287	0,9833	0,0413	0,8779
0,2	0,0518	0,8044	0,0321	0,9486	0,0288	0,9833	0,0288	0,9833	0,0514	0,7470
0,3	0,0681	0,6315	0,0329	0,9295	0,0289	0,9795	0,0289	0,9795	0,0652	0,5969
0,4	0,0873	0,5098	0,0344	0,8947	0,0290	0,9795	0,0290	0,9795	0,0819	0,4368
0,5	0,1156	0,3898	0,0374	0,8241	0,0305	0,9960	0,0305	0,9960	0,1156	0,2939
0,6	0,1569	0,2549	0,0444	0,7472	0,0307	0,9889	0,0305	0,9937	0,1449	0,1802
0,7	0,2242	0,1339	0,0510	0,6004	0,0312	0,9528	0,0319	0,9619	0,2014	0,0997
0,8	0,3003	0,0889	0,0775	0,3958	0,0328	0,8696	0,0337	0,8871	0,3853	0,0633
Média	**0,1306**	**0,4614**	**0,0424**	**0,7888**	**0,0301**	**0,9666**	**0,0302**	**0,9705**	**0,1359**	**0,4120**

Tabela 131 - Precisão e Lembrança para expansão vetorial com *k*=50, janela de 200, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**F**	**F**	**F**	**F**	**F**
0,1	0,0774	0,0579	0,0559	0,0559	0,0789
0,2	0,0972	0,0621	0,0560	0,0559	0,0962
0,3	0,1230	0,0635	0,0562	0,0561	0,1175
0,4	0,1490	0,0663	0,0562	0,0563	0,1379
0,5	0,1783	0,0715	0,0591	0,0591	0,1659
0,6	0,1943	0,0839	0,0596	0,0592	0,1606
0,7	0,1677	0,0939	0,0605	0,0617	0,1334
0,8	0,1372	0,1297	0,0631	0,0649	0,1087
Média	**0,1405**	**0,0786**	**0,0583**	**0,0586**	**0,1249**

Tabela 132 - Medida *F* para expansão vetorial com *k*=50, janela de 200, e limiar de 0.1 até 0.8

Apêndice II – Tabelas de Resultado da Validação do Modelo no contexto de Agrupamento de Documentos

Sem Janela

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,4568	1,9806	2,3647	2,3647	1,6970
0,2	1,7980	2,4914	2,9254	2,9254	2,1101
0,3	3,0798	4,3284	4,9997	4,9997	3,6643
0,4	4,8980	7,1461	7,9504	7,9504	6,1879
0,5	7,5305	11,3105	12,1800	12,1800	10,0869
0,6	11,3829	17,4868	18,2895	18,2895	16,6771
0,7	17,3137	27,5463	27,1765	27,1765	27,2443
0,8	40,0299	62,8360	59,3441	59,3441	69,3669
Média	**10,9362**	**16,8908**	**16,9038**	**16,9038**	**17,1293**

Tabela 133 - Erro quadrático considerando *k*=5, sem janela, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,4038	1,9661	2,3475	2,3475	1,6579
0,2	1,7325	2,4763	2,9050	2,9050	2,0578
0,3	2,9610	4,3090	4,9648	4,9648	3,5753
0,4	4,7180	6,9934	7,8949	7,8949	6,0078
0,5	7,2296	11,1639	12,0950	12,0950	9,6651
0,6	10,9286	17,4372	18,1618	18,1618	15,9518
0,7	16,6331	27,2230	26,9868	26,9868	25,6673
0,8	38,4860	61,1459	58,9299	58,9299	62,2614
Média	**10,5116**	**16,5894**	**16,7857**	**16,7857**	**15,8555**

Tabela 134 - Erro quadrático considerando *k*=10, sem janela, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,3425	1,9298	2,3405	2,3405	1,6113
0,2	1,6565	2,4546	2,8968	2,8968	2,0010
0,3	2,8311	4,3044	4,9508	4,9508	3,4938
0,4	4,5106	6,9461	7,8728	7,8728	5,6735
0,5	6,9122	10,9802	12,0610	12,0610	9,3716
0,6	10,4523	17,0314	18,1108	18,1108	16,3142
0,7	15,8319	25,9925	26,9110	26,9110	24,8223
0,8	36,5840	58,0058	58,7644	58,7644	60,4411
Média	**10,0151**	**15,9556**	**16,7385**	**16,7385**	**15,4661**

Tabela 135 - Erro quadrático considerando *k*=15, sem janela, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,2798	1,8769	2,3402	2,3402	1,5663
0,2	1,5790	2,3820	2,8958	2,8958	1,9571
0,3	2,6987	4,1407	4,9492	4,9492	3,3945
0,4	4,2988	6,6837	7,8702	7,8702	5,6193
0,5	6,5935	10,6272	12,0570	12,0570	9,1501
0,6	9,9241	16,2719	18,1049	18,1049	15,2376
0,7	15,0861	25,6846	26,9021	26,9021	25,3111
0,8	34,6169	58,3088	58,7451	58,7451	60,0894
Média	**9,5096**	**15,7470**	**16,7331**	**16,7331**	**15,2907**

Tabela 136 - Erro quadrático considerando *k*=20, sem janela, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,2062	1,8219	2,3393	2,3393	1,5015
0,2	1,4879	2,3246	2,8956	2,8956	1,8684
0,3	2,5431	4,0491	4,9488	4,9488	3,2444
0,4	4,0522	6,6019	7,8696	7,8696	5,3617
0,5	6,2140	10,4085	12,0561	12,0561	8,8052
0,6	9,3468	16,1181	18,1034	18,1034	14,5294
0,7	14,1915	24,6885	26,9000	26,9000	23,1212
0,8	32,4640	56,0676	58,7405	58,7405	57,5075
Média	**8,9382**	**15,2600**	**16,7317**	**16,7317**	**14,4924**

Tabela 137 - Erro quadrático considerando *k*=25, sem janela, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,1229	1,7825	2,3002	2,3002	1,4502
0,2	1,3850	2,2802	2,8455	2,8455	1,8032
0,3	2,3672	3,9697	4,8631	4,8631	3,1222
0,4	3,7696	6,5225	7,7333	7,7333	5,1683
0,5	5,7844	10,2579	11,8473	11,8473	8,3087
0,6	8,6954	15,7043	17,7899	17,7899	13,6316
0,7	13,1108	24,5624	26,4341	26,4341	22,1436
0,8	29,7780	55,6558	57,7231	57,7231	55,9187
Média	**8,2517**	**15,0919**	**16,4421**	**16,4421**	**13,9433**

Tabela 138 - Erro quadrático considerando *k*=30, sem janela, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,0127	1,7776	2,2405	2,2405	1,3991
0,2	1,2481	2,2473	2,7793	2,7793	1,7500
0,3	2,1341	3,9357	4,7685	4,7685	3,0155
0,4	3,3964	6,4047	7,5829	7,5829	5,0309
0,5	5,2168	10,2311	11,6169	11,6169	8,0755
0,6	7,8574	15,7912	17,4440	17,4440	12,7205
0,7	11,9106	24,4729	25,9201	25,9201	20,8566
0,8	27,0472	54,8573	56,6008	56,6008	54,6019
Média	**7,4779**	**14,9647**	**16,1191**	**16,1191**	**13,4313**

Tabela 139 - Erro quadrático considerando *k*=35, sem janela, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,8757	1,7334	2,1981	2,1981	1,3340
0,2	1,0795	2,1831	2,7239	2,7239	1,6620
0,3	1,8451	3,8921	4,6730	4,6730	2,9078
0,4	2,9350	6,3634	7,4310	7,4310	4,7849
0,5	4,5096	9,9891	11,3987	11,3987	7,4945
0,6	6,7893	15,2601	17,1163	17,1163	11,7674
0,7	10,1927	23,6924	25,4333	25,4333	19,1219
0,8	23,2712	53,3372	55,5376	55,5376	47,3636
Média	**6,4373**	**14,5563**	**15,8140**	**15,8140**	**12,0545**

Tabela 140 - Erro quadrático considerando *k*=40, sem janela, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,6485	1,6131	1,9237	1,9237	1,2202
0,2	0,7995	2,0599	2,3811	2,3811	1,5194
0,3	1,3666	3,6132	4,0707	4,0707	2,6616
0,4	2,1766	5,9946	6,5238	6,5238	4,2943
0,5	3,3440	9,3506	9,9943	9,9943	6,7592
0,6	5,0448	14,2693	15,0269	15,0269	10,7165
0,7	7,5629	22,0851	22,3285	22,3285	17,0739
0,8	17,3658	18,3658	19,3658	20,3658	21,3658
Média	**4,7886**	**9,6689**	**10,2018**	**10,3268**	**8,2014**

Tabela 141 - Erro quadrático considerando *k*=45, sem janela, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,3501	1,3906	1,6632	1,6632	1,0612
0,2	0,4316	1,7384	2,0731	2,0731	1,3274
0,3	0,7384	3,0563	3,5748	3,5748	2,2809
0,4	1,1931	4,9776	5,7120	5,7120	3,6902
0,5	1,8650	7,8226	8,7898	8,7898	5,8063
0,6	2,8720	11,8885	13,3103	13,3103	9,0940
0,7	4,5552	18,5720	19,8112	19,8112	14,5088
0,8	12,2116	41,6527	43,5866	43,5866	37,8335
Média	**3,0271**	**11,3873**	**12,3151**	**12,3151**	**9,4503**

Tabela 142 - Erro quadrático considerando *k*=50, sem janela, e limiar de 0.1 até 0.8

Janela de 20

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	2,2579	2,2919	2,4592	2,3568	2,2956
0,2	2,7829	2,8337	3,0399	2,9112	2,8362
0,3	4,7577	4,8430	5,1954	5,0073	4,8608
0,4	7,5740	7,7014	8,2618	7,9850	7,7452
0,5	11,6061	11,7984	12,6570	12,2915	11,9289
0,6	17,6709	17,7166	19,0057	18,8390	18,0974
0,7	26,7154	26,3251	28,2407	28,1467	27,1115
0,8	60,9781	57,4852	61,6681	61,5973	60,4953
Média	**16,7929**	**16,3744**	**17,5660**	**17,3918**	**16,9214**

Tabela 143 - Erro quadrático considerando *k*=5, janela de 20, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	2,1655	2,2834	2,4528	2,3113	2,2414
0,2	2,6695	2,8228	3,0318	2,8925	2,7712
0,3	4,5639	4,8243	5,1816	4,9621	4,7705
0,4	7,2635	7,6716	8,2398	7,9687	7,5881
0,5	11,1370	11,7528	12,6233	12,2169	11,6865
0,6	16,8431	17,6480	18,9551	18,4667	17,6276
0,7	25,6330	26,2233	28,1655	27,7212	26,4872
0,8	58,6020	57,2627	61,5039	60,8231	59,1814
Média	**16,1097**	**16,3111**	**17,5192**	**17,1703**	**16,5442**

Tabela 144 - Erro quadrático considerando *k*=10, janela de 20, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	2,0599	2,2555	2,4482	2,2414	2,1890
0,2	2,5390	2,8052	3,0259	2,8467	2,7064
0,3	4,3407	4,7943	5,1714	4,7989	4,6446
0,4	6,9096	7,6239	8,2236	7,7045	7,3952
0,5	10,5890	11,6797	12,5985	12,0320	11,3889
0,6	16,1633	17,5382	18,9179	18,1283	17,1893
0,7	24,2618	26,0602	28,1103	27,1202	25,8309
0,8	55,6508	56,9065	61,3833	60,1314	57,7450
Média	**15,3143**	**16,2079**	**17,4849**	**16,8754**	**16,1362**

Tabela 145 - Erro quadrático considerando *k*=15, janela de 20, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,9541	2,2666	2,4369	2,2582	2,1076
0,2	2,4084	2,8025	3,0119	2,7846	2,6174
0,3	4,1220	4,7896	5,1476	4,7970	4,4698
0,4	6,5631	7,6164	8,1857	7,7533	7,1470
0,5	10,0591	11,6682	12,5404	11,9763	11,0459
0,6	15,2242	17,5210	18,8308	18,0699	16,6838
0,7	23,0773	26,0346	27,9807	27,1815	25,1314
0,8	52,6701	56,8506	61,1004	60,0651	55,7659
Média	**14,5098**	**16,1937**	**17,4043**	**16,8607**	**15,6211**

Tabela 146 - Erro quadrático considerando *k*=20, janela de 20, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,8338	2,2613	2,4199	2,1973	2,0420
0,2	2,2604	2,7960	2,9950	2,7401	2,5291
0,3	3,8672	4,7785	5,1186	4,7686	4,3420
0,4	6,1527	7,5988	8,1396	7,6175	6,9546
0,5	9,4351	11,6413	12,4697	11,7021	10,7615
0,6	14,2428	17,4806	18,7246	17,7534	16,3310
0,7	21,4347	25,9746	27,8229	26,6195	24,7443
0,8	49,6270	56,7197	60,7558	58,7416	54,8484
Média	**13,6067**	**16,1564**	**17,3058**	**16,5175**	**15,3191**

Tabela 147 - Erro quadrático considerando *k*=25, janela de 20, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,6861	2,2468	2,3874	2,1300	1,9381
0,2	2,0780	2,7786	2,9519	2,6428	2,4322
0,3	3,5549	4,7488	5,0451	4,5719	4,1648
0,4	5,6616	7,5515	8,0227	7,3890	6,6805
0,5	8,6801	11,5689	12,2906	11,4091	10,3544
0,6	13,0518	17,3718	18,4557	17,3224	15,8272
0,7	19,6363	25,8129	27,4234	25,8085	23,6089
0,8	45,2027	56,3667	59,8833	57,2150	52,9657
Média	**12,4439**	**16,0558**	**17,0575**	**16,0611**	**14,7465**

Tabela 148 - Erro quadrático considerando *k*=30, janela de 20, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,5048	2,1995	2,3303	2,0365	1,8247
0,2	1,8547	2,7106	2,9022	2,5498	2,2679
0,3	3,1721	4,6463	4,9601	4,4638	3,9112
0,4	5,0548	7,3886	7,8875	7,1248	6,2992
0,5	7,7439	11,3325	12,0835	11,0018	9,7263
0,6	11,6612	17,0169	18,1447	16,4781	14,7689
0,7	17,5248	25,2856	26,9613	24,7947	22,3531
0,8	40,4200	55,2151	58,8743	55,3262	49,9623
Média	**11,1170**	**15,7244**	**16,7680**	**15,4720**	**13,8892**

Tabela 149 - Erro quadrático considerando *k*=35, janela de 20, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,2736	2,1675	2,2337	1,9663	1,6910
0,2	1,5696	2,6816	2,7734	2,4613	2,1082
0,3	2,6878	4,5990	4,7561	4,2608	3,6129
0,4	4,2742	7,3133	7,5631	6,8131	5,8388
0,5	6,5527	11,2038	11,5866	10,6334	9,0717
0,6	9,8603	16,8384	17,3985	16,2026	13,9929
0,7	14,8991	25,0203	25,8525	24,1296	20,8248
0,8	34,1502	54,6358	56,5076	53,2716	46,4569
Média	**9,4084**	**15,5575**	**16,0839**	**14,9673**	**12,9496**

Tabela 150 - Erro quadrático considerando *k*=40, janela de 20, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,9525	1,8775	1,9631	1,8086	1,4133
0,2	1,1741	2,3229	2,4326	2,2828	1,7669
0,3	2,0089	3,9700	4,1576	3,9879	3,1568
0,4	3,2005	6,3143	6,6400	6,4033	5,0520
0,5	4,9090	9,7686	10,1721	9,8610	7,7980
0,6	7,3764	14,6685	15,2565	14,9936	11,7358
0,7	11,2089	21,7961	22,6697	22,4047	17,9804
0,8	25,1342	47,6957	49,9100	49,3850	41,0506
Média	**6,9956**	**13,5517**	**14,1502**	**13,8909**	**11,2442**

Tabela 151 - Erro quadrático considerando *k*=45, janela de 20, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,5014	1,6525	1,6570	1,6557	1,1293
0,2	0,6180	2,0661	2,0658	2,0731	1,4349
0,3	1,0583	3,5537	3,5545	3,5768	2,5055
0,4	1,6850	5,6720	5,6647	5,7523	3,9903
0,5	2,6472	8,7394	8,7479	9,0369	6,1977
0,6	4,0403	13,1674	13,1845	13,6416	9,3319
0,7	6,4105	19,9160	19,7576	20,5972	14,4091
0,8	15,5243	43,3273	43,1078	45,6054	32,3260
Média	**4,0606**	**12,2618**	**12,2175**	**12,7424**	**8,9156**

Tabela 152 - Erro quadrático considerando *k*=50, janela de 20, e limiar de 0.1 até 0.8

Janela de 50

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,8012	2,3325	2,4874	2,4216	2,4351
0,2	2,2201	2,8862	3,0769	3,0235	3,0070
0,3	3,7999	4,9327	5,2586	5,2652	5,1550
0,4	6,0479	7,8439	8,3622	8,3698	8,2135
0,5	9,2809	12,0168	12,8108	12,9103	12,6305
0,6	14,0780	18,0444	19,2368	19,4982	19,0190
0,7	21,5554	26,8124	28,5840	29,2881	28,5184
0,8	49,2975	58,5491	62,4178	64,1815	63,8709
Média	**13,5101**	**16,6772**	**17,7793**	**18,1198**	**17,8562**

Tabela 153 - Erro quadrático considerando *k*=5, janela de 50, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,7247	2,3224	2,4744	2,3507	2,3733
0,2	2,1259	2,8748	3,0613	2,9259	2,9353
0,3	3,6391	4,9132	5,2320	5,0791	5,0296
0,4	5,7919	7,8130	8,3200	8,1651	8,0243
0,5	8,8923	11,9695	12,7461	12,6269	12,3130
0,6	13,4576	17,9734	19,1396	19,0210	18,5471
0,7	20,5619	26,7067	28,4396	28,7280	27,7819
0,8	47,4058	58,3184	62,1024	62,7829	61,8399
Média	**12,9499**	**16,6114**	**17,6894**	**17,7099**	**17,3555**

Tabela 154 - Erro quadrático considerando *k*=10, janela de 50, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,6522	2,3217	2,4662	2,3517	2,3076
0,2	2,0364	2,8736	3,0519	2,9382	2,8546
0,3	3,4859	4,9111	5,2160	5,0524	4,8934
0,4	5,5480	7,8096	8,2944	8,1632	7,8034
0,5	8,5102	11,9642	12,7069	12,5748	12,0068
0,6	12,8840	17,9655	19,0807	18,8746	18,0692
0,7	19,7315	26,6951	28,3522	28,3606	27,1339
0,8	45,2445	58,2929	61,9115	62,8666	60,4251
Média	**12,3866**	**16,6042**	**17,6350**	**17,6478**	**16,9368**

Tabela 155 - Erro quadrático considerando *k*=15, janela de 50, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,5634	2,3137	2,4633	2,2779	2,2342
0,2	1,9273	2,8643	3,0488	2,8475	2,7597
0,3	3,2986	4,8953	5,2106	4,9349	4,7391
0,4	5,2473	7,7845	8,2859	7,9889	7,5736
0,5	8,0508	11,9257	12,6940	12,3973	11,6178
0,6	12,1923	17,9077	19,0613	18,6015	17,5040
0,7	18,6061	26,6091	28,3232	27,8003	26,3170
0,8	42,7399	58,1052	61,8483	61,3816	58,5213
Média	**11,7032**	**16,5507**	**17,6169**	**17,2787**	**16,4083**

Tabela 156 - Erro quadrático considerando *k*=20, janela de 50, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,4714	2,2865	2,4506	2,2640	2,1543
0,2	1,8139	2,8294	3,0310	2,8441	2,6588
0,3	3,1017	4,8356	5,1801	4,9371	4,5773
0,4	4,9462	7,6895	8,2375	7,9716	7,3258
0,5	7,5926	11,7802	12,6197	12,2139	11,2714
0,6	11,4621	17,6892	18,9498	18,5908	16,9186
0,7	17,3463	26,2845	28,1576	27,7471	25,5491
0,8	40,0646	57,3963	61,4865	61,3559	56,6633
Média	**10,9749**	**16,3489**	**17,5141**	**17,2406**	**15,8898**

Tabela 157 - Erro quadrático considerando *k*=25, janela de 50, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,3669	2,2644	2,3850	2,1858	2,0418
0,2	1,6850	2,8012	2,9507	2,7606	2,5280
0,3	2,8830	4,7874	5,0430	4,7955	4,3489
0,4	4,5930	7,6128	8,0193	7,7398	6,9307
0,5	7,0444	11,6628	12,2855	11,9767	10,6471
0,6	10,6588	17,5128	18,4479	18,2419	16,0039
0,7	16,0918	26,0225	27,4118	27,2202	24,0571
0,8	37,0424	56,8242	59,8581	60,1463	53,5982
Média	**10,1707**	**16,1860**	**17,0501**	**16,8833**	**15,0195**

Tabela 158 - Erro quadrático considerando *k*=30, janela de 50, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,2232	2,2324	2,3407	2,1104	1,9334
0,2	1,5077	2,7621	2,9044	2,6582	2,3959
0,3	2,5797	4,7392	4,9835	4,6481	4,1168
0,4	4,1046	7,5363	7,9247	7,4128	6,5481
0,5	6,3023	11,5456	12,1406	11,5236	10,0849
0,6	9,5149	17,3369	18,2304	17,7703	15,1879
0,7	14,3568	25,7610	27,0886	26,3015	22,8642
0,8	32,7874	56,2532	59,1524	58,1697	50,4490
Média	**9,0471**	**16,0209**	**16,8457**	**16,3243**	**14,1975**

Tabela 159 - Erro quadrático considerando *k*=35, janela de 50, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,0453	2,1780	2,3011	2,0045	1,7911
0,2	1,2883	2,7092	2,8549	2,4971	2,2187
0,3	2,2022	4,6303	4,8977	4,4313	3,8015
0,4	3,5066	7,3630	7,7883	7,1188	6,0742
0,5	5,3795	11,2950	11,9479	11,0525	9,3448
0,6	8,1020	16,9606	17,9410	16,8311	14,1263
0,7	12,4442	25,2019	26,6587	25,3738	21,3230
0,8	28,2498	55,0325	58,2134	55,1201	46,8478
Média	**7,7772**	**15,6713**	**16,5754**	**15,5536**	**13,1909**

Tabela 160 - Erro quadrático considerando *k*=40, janela de 50, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,7772	1,9055	2,0052	1,8309	1,5210
0,2	0,9583	2,3625	2,4939	2,2933	1,8894
0,3	1,6394	4,0383	4,2623	4,0786	3,2634
0,4	2,6090	6,4237	6,7530	6,5293	5,1969
0,5	4,0062	9,9260	10,3898	10,1662	8,0716
0,6	6,0534	14,9049	15,6014	15,3399	12,2081
0,7	9,0867	22,1941	23,2298	22,7300	18,3289
0,8	20,4659	48,5341	50,8088	50,2980	40,6270
Média	**5,6995**	**13,7861**	**14,4430**	**14,1583**	**11,3883**

Tabela 161 - Erro quadrático considerando *k*=45, janela de 50, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,4131	1,6590	1,6615	1,2975	1,0926
0,2	0,5091	2,0661	2,0821	1,6402	1,3522
0,3	0,8716	3,5628	3,5905	2,8953	2,3401
0,4	1,4043	5,6895	5,7221	4,5667	3,7564
0,5	2,1889	8,7533	8,8085	7,2010	5,7997
0,6	3,4264	13,2060	13,3919	10,9712	8,9054
0,7	5,3166	19,9110	19,9419	16,6857	13,2941
0,8	13,3415	43,3925	43,5043	36,7848	29,1289
Média	**3,4339**	**12,2800**	**12,3379**	**10,2553**	**8,2087**

Tabela 162 - Erro quadrático considerando *k*=50, janela de 50, e limiar de 0.1 até 0.8

Janela de 100

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,6035	1,4734	2,4953	2,3541	2,2961
0,2	1,9789	1,8295	3,0881	2,9132	2,8353
0,3	3,3901	3,1338	5,2778	4,9788	4,8565
0,4	5,3991	5,0171	8,3927	7,9173	7,7421
0,5	8,2817	7,7245	12,8576	12,1291	11,9078
0,6	12,5249	11,5651	19,3069	18,2131	17,9871
0,7	19,1357	17,2009	28,6883	27,0630	26,9652
0,8	44,2569	37,7700	62,6454	59,0965	59,7061
Média	**12,0714**	**10,7143**	**17,8440**	**16,8331**	**16,7870**

Tabela 163 - Erro quadrático considerando *k*=5, janela de 100, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,5379	1,4487	2,4878	2,3377	2,2622
0,2	1,8976	1,8025	3,0794	2,8937	2,7919
0,3	3,2510	3,0829	5,2629	4,9454	4,7794
0,4	5,1781	4,9281	8,3691	7,8642	7,6292
0,5	7,9387	7,5588	12,8213	12,0479	11,7426
0,6	12,0166	11,3741	19,2525	18,0912	17,7168
0,7	18,2443	16,9097	28,6074	26,8818	26,7288
0,8	42,2077	37,0458	62,4688	58,7006	59,1333
Média	**11,5340**	**10,5188**	**17,7936**	**16,7203**	**16,5980**

Tabela 164 - Erro quadrático considerando *k*=10, janela de 100, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,4727	1,4406	2,4757	2,3408	2,1903
0,2	1,8172	1,7834	3,0655	2,8935	2,7044
0,3	3,1122	3,0652	5,2392	4,9452	4,6314
0,4	4,9532	4,8889	8,3314	7,8638	7,3928
0,5	7,5981	7,5221	12,7636	12,0472	11,3824
0,6	11,4714	11,2572	19,1659	18,0902	17,2074
0,7	17,3635	16,5968	28,4787	26,8803	25,9199
0,8	40,7631	36,6657	62,1877	58,6974	57,3881
Média	**11,0689**	**10,4025**	**17,7135**	**16,7198**	**16,1021**

Tabela 165 - Erro quadrático considerando *k*=15, janela de 100, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,3985	1,4127	2,4756	2,3367	2,1363
0,2	1,7251	1,7598	3,0651	2,8926	2,6364
0,3	2,9539	3,0211	5,2385	4,9436	4,5207
0,4	4,6992	4,8308	8,3303	7,8613	7,2207
0,5	7,2062	7,4319	12,7619	12,0434	11,1145
0,6	10,8804	11,1480	19,1633	18,0844	16,7213
0,7	16,6094	16,3654	28,4749	26,8718	25,2258
0,8	38,2844	36,0825	62,1794	58,6788	55,9261
Média	**10,4696**	**10,2565**	**17,7111**	**16,7141**	**15,6877**

Tabela 166 - Erro quadrático considerando *k*=20, janela de 100, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,3174	1,3985	2,4585	2,3351	2,0711
0,2	1,6250	1,7388	3,0420	2,8898	2,5606
0,3	2,7775	2,9802	5,1990	4,9388	4,3886
0,4	4,4278	4,7545	8,2674	7,8537	7,0083
0,5	6,7844	7,2967	12,6655	12,0318	10,7497
0,6	10,2270	10,9640	19,0186	18,0669	16,2359
0,7	15,5109	16,2967	28,2598	26,8457	24,4265
0,8	35,7632	35,7404	61,7098	58,6219	54,2815
Média	**9,8042**	**10,1462**	**17,5776**	**16,6980**	**15,2153**

Tabela 167 - Erro quadrático considerando *k*=25, janela de 100, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,2206	1,3733	2,4006	2,2950	1,9745
0,2	1,5051	1,7043	2,9707	2,8392	2,4392
0,3	2,5726	2,9214	5,0771	4,8523	4,1915
0,4	4,0951	4,6620	8,0735	7,7162	6,6800
0,5	6,2892	7,1551	12,3686	11,8211	10,2898
0,6	9,4537	10,7550	18,5727	17,7505	15,5097
0,7	14,3483	16,0137	27,5973	26,3756	23,3278
0,8	32,5232	35,1071	60,2631	57,5955	51,7260
Média	**9,0010**	**9,9615**	**17,1654**	**16,4057**	**14,5173**

Tabela 168 - Erro quadrático considerando *k*=30, janela de 100, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,1002	1,3689	2,3481	2,2366	1,8501
0,2	1,3566	1,7038	2,9142	2,7736	2,2867
0,3	2,3188	2,9067	5,0017	4,7595	3,9230
0,4	3,6922	4,6654	7,9537	7,5686	6,2615
0,5	5,6692	7,1480	12,1850	11,5950	9,6442
0,6	8,5481	10,7213	18,2971	17,4111	14,5736
0,7	12,8705	15,8965	27,1877	25,8713	21,9466
0,8	28,8688	34,7368	59,3688	56,4940	48,4762
Média	**8,0530**	**9,8934**	**16,9071**	**16,0887**	**13,6202**

Tabela 169 - Erro quadrático considerando *k*=35, janela de 100, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,9395	1,3192	2,3128	2,2008	1,7114
0,2	1,1579	1,6450	2,8692	2,7175	2,1200
0,3	1,9801	2,8344	4,9234	4,6615	3,6377
0,4	3,1501	4,5365	7,8291	7,4127	5,7895
0,5	4,8473	6,9497	12,0109	11,3711	8,9333
0,6	7,3059	10,4457	18,0356	17,0749	13,4956
0,7	11,1050	15,4816	26,7992	25,3716	20,3136
0,8	25,0817	34,0555	58,5203	55,4030	44,8546
Média	**6,9459**	**9,6584**	**16,6625**	**15,7766**	**12,6070**

Tabela 170 - Erro quadrático considerando *k*=40, janela de 100, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,7017	1,2824	2,0138	1,9196	1,4316
0,2	0,8649	1,6064	2,5093	2,3745	1,7797
0,3	1,4785	2,7525	4,2885	4,0587	3,0486
0,4	2,3521	4,3910	6,7948	6,4563	4,8705
0,5	3,6144	6,7838	10,4538	9,9757	7,5227
0,6	5,4585	10,1799	15,6974	14,9795	11,3660
0,7	8,1894	15,0591	23,3747	22,3038	17,0503
0,8	18,9270	32,7914	51,1304	48,7762	37,9936
Média	**5,1983**	**9,3558**	**14,5328**	**13,8555**	**10,6329**

Tabela 171 - Erro quadrático considerando *k*=45, janela de 100, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,3726	1,0175	1,6584	1,6630	0,9850
0,2	0,4592	1,2807	2,0872	2,0720	1,2222
0,3	0,7856	2,2223	3,5933	3,5716	2,0940
0,4	1,2678	3,5500	5,7333	5,6972	3,3578
0,5	1,9816	5,4166	8,8335	8,7733	5,1811
0,6	3,0952	8,2216	13,4539	13,2646	7,8088
0,7	4,8450	12,2637	20,0029	19,8825	11,6377
0,8	12,7152	27,5951	43,6699	43,5296	25,6531
Média	**3,1903**	**7,6959**	**12,3790**	**12,3067**	**7,2424**

Tabela 172 - Erro quadrático considerando *k*=50, janela de 100, e limiar de 0.1 até 0.8

Janela de 200

k=5	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,4949	1,7130	2,5014	2,3600	2,0823
0,2	1,8449	2,1754	3,0962	2,9201	2,5731
0,3	3,1603	3,8073	5,2916	4,9906	4,4086
0,4	5,0295	6,1209	8,4147	7,9361	7,0322
0,5	7,7175	9,3771	12,8912	12,1580	10,8203
0,6	11,6847	14,0806	19,3574	18,2564	16,2648
0,7	17,7138	20,9225	28,7633	27,1273	24,4013
0,8	41,2039	45,6877	62,8092	59,2368	53,9936
Média	**11,2312**	**12,9856**	**17,8906**	**16,8732**	**15,1970**

Tabela 173 - Erro quadrático considerando *k*=5, janela de 200, e limiar de 0.1 até 0.8

k=10	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,4376	1,6779	2,4919	2,3426	2,0205
0,2	1,7741	2,1334	3,0854	2,8994	2,4962
0,3	3,0377	3,7404	5,2732	4,9552	4,2759
0,4	4,8317	5,9896	8,3854	7,8797	6,8216
0,5	7,4111	9,1822	12,8463	12,0717	10,4957
0,6	11,2059	13,7881	19,2901	18,1269	15,8304
0,7	17,0580	20,4878	28,6632	26,9348	23,7319
0,8	39,2399	44,7383	62,5907	58,8165	52,2553
Média	**10,7495**	**12,7172**	**17,8283**	**16,7534**	**14,7409**

Tabela 174 - Erro quadrático considerando *k*=10, janela de 200, e limiar de 0.1 até 0.8

k=15	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,3748	1,5817	2,4811	2,3407	1,9440
0,2	1,6964	2,0545	3,0729	2,8971	2,4008
0,3	2,9044	3,5145	5,2517	4,9513	4,1172
0,4	4,6201	5,6511	8,3513	7,8735	6,5715
0,5	7,0878	8,6575	12,7941	12,0621	10,1034
0,6	10,7036	13,0001	19,2116	18,1125	15,2639
0,7	16,2821	19,3169	28,5466	26,9135	22,8920
0,8	37,5340	42,1816	62,3362	58,7698	50,5139
Média	**10,2754**	**11,9947**	**17,7557**	**16,7401**	**14,2258**

Tabela 175 - Erro quadrático considerando *k*=15, janela de 200, e limiar de 0.1 até 0.8

k=20	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,3085	1,5505	2,4804	2,3403	1,8403
0,2	1,6146	1,9592	3,0717	2,8960	2,2767
0,3	2,7595	3,3718	5,2497	4,9494	3,9028
0,4	4,3963	5,3911	8,3480	7,8705	6,2103
0,5	6,7482	8,4407	12,7891	12,0576	9,5829
0,6	10,1514	12,6746	19,2041	18,1057	14,4372
0,7	15,4869	18,8333	28,5354	26,9033	21,6853
0,8	35,6663	41,1256	62,3117	58,7477	48,0902
Média	**9,7665**	**11,6684**	**17,7488**	**16,7338**	**13,5032**

Tabela 176 - Erro quadrático considerando *k*=20, janela de 200, e limiar de 0.1 até 0.8

k=25	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,2370	1,4875	2,4589	2,3394	1,7928
0,2	1,5259	1,9403	3,0429	2,8958	2,2171
0,3	2,6086	3,3546	5,2004	4,9491	3,7994
0,4	4,1487	5,3368	8,2697	7,8701	6,0583
0,5	6,3717	8,1760	12,6691	12,0569	9,3192
0,6	9,5903	12,2771	19,0240	18,1047	14,0550
0,7	14,5343	18,2426	28,2678	26,9018	21,0757
0,8	33,3236	39,8356	61,7274	58,7444	46,8962
Média	**9,1675**	**11,3313**	**17,5825**	**16,7328**	**13,1517**

Tabela 177 - Erro quadrático considerando *k*=25, janela de 200, e limiar de 0.1 até 0.8

k=30	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,1475	1,4745	2,3975	2,2992	1,6935
0,2	1,4154	1,8288	2,9673	2,8443	2,0919
0,3	2,4191	3,2795	5,0714	4,8611	3,5884
0,4	3,8499	5,2335	8,0645	7,7301	5,7249
0,5	5,9127	8,0176	12,3547	11,8424	8,8043
0,6	8,8882	12,1858	18,5519	17,7825	13,2928
0,7	13,4562	18,1675	27,5663	26,4232	19,9543
0,8	30,5490	39,6716	60,1955	57,6993	44,3154
Média	**8,4548**	**11,2323**	**17,1461**	**16,4353**	**12,4332**

Tabela 178 - Erro quadrático considerando *k*=30, janela de 200, e limiar de 0.1 até 0.8

k=35	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	1,0356	1,4388	2,3667	2,2404	1,5846
0,2	1,2763	1,8705	2,9167	2,7789	1,9600
0,3	2,1825	3,2269	5,0066	4,7684	3,3719
0,4	3,4732	5,1314	7,9615	7,5827	5,3875
0,5	5,3357	7,8558	12,1969	11,6166	8,2887
0,6	8,0352	11,7963	18,3149	17,4435	12,4818
0,7	12,1940	17,5282	27,2142	25,9194	18,7903
0,8	27,2950	38,2756	59,4266	56,5992	41,5550
Média	**7,6034**	**10,8904**	**16,9255**	**16,1186**	**11,6775**

Tabela 179 - Erro quadrático considerando *k*=35, janela de 200, e limiar de 0.1 até 0.8

k=40	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,8931	1,4534	2,3352	2,1980	1,4455
0,2	1,1010	1,8410	2,8778	2,7239	1,7931
0,3	1,8819	3,1829	4,9339	4,6729	3,0800
0,4	2,9935	5,0614	7,8459	7,4308	4,9172
0,5	4,6062	7,7490	12,0369	11,3985	7,5743
0,6	6,9417	11,6359	18,0746	17,1161	11,4830
0,7	10,3723	17,2898	26,8571	25,4329	17,2751
0,8	23,7369	37,7551	58,6468	55,5367	37,9976
Média	**6,5658**	**10,7461**	**16,7010**	**15,8137**	**10,6957**

Tabela 180 - Erro quadrático considerando *k*=40, janela de 200, e limiar de 0.1 até 0.8

k=45	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,6628	1,4565	2,0148	1,9234	1,1839
0,2	0,8169	1,8363	2,5125	2,3807	1,4698
0,3	1,3965	3,1731	4,2940	4,0700	2,5252
0,4	2,2210	5,0458	6,8310	6,5226	4,0406
0,5	3,4134	7,7246	10,4650	10,0054	6,2430
0,6	5,1514	11,5992	15,7142	15,0241	9,4353
0,7	7,7531	17,2354	23,4012	22,3245	14,1906
0,8	18,0848	37,6362	51,1704	48,8440	31,3408
Média	**4,9375**	**10,7134**	**14,5504**	**13,8868**	**8,8036**

Tabela 181 - Erro quadrático considerando *k*=45, janela de 200, e limiar de 0.1 até 0.8

k=50	LRD	Phi-squared	MI	VMI	Z score
Limiar	**Erro**	**Erro**	**Erro**	**Erro**	**Erro**
0,1	0,3554	0,8947	1,6654	1,6630	0,8226
0,2	0,4381	1,1904	2,0880	2,0728	1,0172
0,3	0,7496	2,2333	3,5961	3,5743	1,7436
0,4	1,2106	3,5770	5,7395	5,7113	2,8054
0,5	1,8922	5,6904	8,8373	8,7889	4,3178
0,6	2,8895	9,3325	13,4519	13,3087	6,5186
0,7	4,6206	14,0854	20,0084	19,8086	9,7460
0,8	12,3192	30,7577	43,6631	43,5968	21,3502
Média	**3,0594**	**8,4702**	**12,3812**	**12,3156**	**6,0402**

Tabela 182 - Erro quadrático considerando *k*=50, janela de 200, e limiar de 0.1 até 0.8