Gleidson Antônio Cardoso da Silva

KDC: UMA ABORDAGEM BASEADA EM CONHECIMENTO PARA CLASSIFICAÇÃO DE DOCUMENTOS

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do Grau de Mestre em Ciência da Computação.
Orientadora: Profª. Drª. Carina Friedrich Dorneles

Florianópolis
2015

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

da Silva, Gleidson Antônio Cardoso
KDC: Uma Abordagem Baseada em Conhecimento para Classificação de Documentos / Gleidson Antônio Cardoso da Silva ; orientadora, Carina Friedrich Dorneles - Florianópolis, SC, 2015.
66 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Ciência da Computação.

Inclui referências

1. Ciência da Computação. 2. Classificação de Documentos. 3. Ranqueamento de Classes. 4. Base de Conhecimento. I. Dorneles, Carina Friedrich. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Gleidson Antônio Cardoso da Silva

KDC: UMA ABORDAGEM BASEADA EM CONHECIMENTO PARA CLASSIFICAÇÃO DE DOCUMENTOS

Esta Dissertação foi julgada adequada para obtenção do Título de Mestre e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

Florianópolis, 02 de dezembro de 2015.

Profª. Carina Friedrich Dorneles, Drª.
Coordenadora do Programa

Banca Examinadora:

Profª. Carina Friedrich Dorneles, Drª.
Orientadora
Universidade Federal de Santa Catarina

Prof. Alexandre Leopoldo Gonçalves, Dr.
Universidade Federal de Santa Catarina

Prof. Ronaldo dos Santos Mello, Dr.
Universidade Federal de Santa Catarina

Prof. Frank Augusto Siqueira, Dr.
Universidade Federal de Santa Catarina

Dedicado a Cassiana Vanessa Teixeira
e aos meus pais.

AGRADECIMENTOS

Meus mais sinceros agradecimentos a minha orientadora Profª. Drª. Carina Friedrich Dorneles pela oportunidade, por confiar em meu trabalho e me guiar durante o curso. A minha noiva Cassiana Vanessa Teixeira por todo o apoio, lealdade, compreensão e incentivo. A Gunther Schmitz Lardizabal, pela confiança e por tornar possível minha participação no curso.

RESUMO

Classificação de documentos fornece um meio para organizar as informações, permitindo uma melhor compreensão e interpretação dos dados. A tarefa de classificar é caracterizada pela associação de rótulos de classes a documentos com o objetivo de criar agrupamentos semânticos. O aumento exponencial no número de documentos e dados digitais demanda formas mais precisas, abrangentes e eficientes para busca e organização de informações. Nesse contexto, o aprimoramento de técnicas de classificação de documentos com o uso de informação semântica é considerado essencial. Sendo assim, este trabalho propõe uma abordagem baseada em conhecimento para a classificação de documentos. A técnica utiliza termos extraídos de documentos associando-os a conceitos de uma base de conhecimento de domínio aberto. Em seguida, os conceitos são generalizados a um nível maior de abstração. Por fim, é calculado um valor de disparidade entre os conceitos generalizados e o documento, sendo o conceito de menor disparidade considerado como rótulo de classe aplicável ao documento. A aplicação da técnica proposta oferece vantagens sobre os métodos convencionais como a ausência da necessidade de treinamento, a oportunidade de atribuir uma ou múltiplas classes a um documento e a capacidade de aplicação em diferentes temas de classificação sem a necessidade de alterar o classificador.

Palavras-chave: Classificação de Documentos. Ranqueamento de Classes. Base de Conhecimento.

ABSTRACT

Document classification provides a way to organize information, providing a better way to understand available data. The classification task is characterized by the association of class labels to documents, aiming to create semantic clusters. The exponential increase in the number of documents and digital data demands for more precise, comprehensive and efficient ways to search and organize information. In this context, the improvement of document classification techniques using semantic information is considered essential. Thus, this paper proposes a knowledge-based approach for the classification of documents. The technique uses terms extracted from documents in association with concepts of an open domain knowledge base. Then, the concepts are generalized to a higher level of abstraction. Finally a disparity value between generalized concepts and the document is calculated, and the best ranked concept is then considered as a class label applicable to the document. The application of the proposed technique offers advantages over conventional methods including no need for training, the choice to assign one or multiple classes to a document and the capacity to classify over different subjects without the need to change the classifier.

Keywords: Document Classification. Class Ranking. Knowledge Base.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LSI - Latent Semantic Indexing
KDC - Knowledge-based Document Classification
ML – Machine Learning
RDF- Resource Description Framework
NLP - Natural Language Processing
KNN – K-Nearest Neighbors
DMOZ – The Open Directory Project
TF-IDF - Term Frequency–Inverse Document Frequency
SVM – Support Vector Machine
NB – Naive Bayes
J48 – Decision Trees

SUMÁRIO

1 INTRODUÇÃO

Classificação ou categorização de documentos fornece um meio para organizar as informações, o que permite uma melhor compreensão e interpretação dos dados (YATES; NETO, 2011). Para a área de recuperação de informação, a tarefa de classificação de documentos em múltiplas classes é crucial para muitas aplicações, tais como o desenvolvimento de diretórios Web (DUMAIS; CHEN, 2000), integração empresarial (HALEVY; ASHISH; BITTON; CAREY; DRAPER; POLLOCK; SIKKA, 2005; HE; DA XU, 2014), *Focused Crawling* (VIDAL; DA SILVA; DE MOURA; CAVALCANTI, 2008), e muitas outras (XIAOGUANG; DAVISON, 2009; HERNÁNDEZ; RIVERO; RUIZ; CORCHUELO, 2014; HALL; FRANK; HOLMES; PFAHRINGER; REUTEMANN; WITTEN, 2009; AGGARWAL; ZHAI, 2012). Além disso, a classificação de documentos também pode ser útil na organização da grande quantidade de informação disponível na Internet, para retorno pelos motores de busca, por exemplo.

Nesses cenários, a classificação é o problema da atribuição de pelo menos uma classe ao documento, com base em seu conteúdo. Este problema tem sido amplamente pesquisado na comunidade (DUMAIS; CHEN, 2000; XIAOGUANG; DAVISON, 2009; SIGOGNE; CONSTANT, 2009; LUCIA; FERRARI, 2014; KORDE; MAHENDER, 2012; AGGARWAL; ZHAI, 2012; HUSBY; STEPHANIE; BARBOSA, 2012; TAO; LI; LAU; WANG, 2012; ALLAHYARI; KOCHUT; JANIK, 2014; KO; SEO, 2000; PANG; JIANG, 2013; MATERNA, 2008), e várias técnicas têm sido propostas, incluindo abordagens que usam *Machine Learning* (ML) (HUSBY; STEPHANIE; BARBOSA, 2012; KO; SEO, 2000) e *Natural Language Processing* (NLP) (MATERNA, 2008).

A maioria das abordagens propostas que obtiveram sucesso na tarefa de classificação de documentos adota ML (KORDE; MAHENDER, 2012; AGGARWAL; ZHAI, 2012). Nessas abordagens, conjuntos de palavras ou padrões são usados como modelo de representação de documentos, com a finalidade de treinar um modelo a ser usado para classificação de um conjunto de documentos.

Algumas situações tornam a aplicação de ML mais difícil, e por vezes inviável, como por exemplo:

i. quando não existem documentos suficientes para a formação de um modelo de classificação;
ii. quando a classificação precisa ser realizada sem um conjunto de documentos que possibilite o treinamento de um modelo de classificação;

iii. quando o conjunto de documentos a ser classificado possui tamanho indefinido;
iv. quando é necessário trabalhar com vocabulário crescente e irrestrito.

Uma possível solução para estes problemas com o uso de ML envolve a atualização contínua do modelo utilizado pelo classificador e a validação do modelo de classificação em tempo de execução, aumentando o custo de classificar novos casos, uma vez que a maior parte do algoritmo de cálculo é realizada na etapa de construção do modelo de classificação.

Outro problema refere-se à necessidade de interação humana para classificar manualmente documentos a serem usados pelo conjunto de treinamento. Além disso, a definição da quantidade e tipo de documento que devem ser usados no conjunto de treinamento não é um processo simples (SHEN et. al, 2004). Essas definições devem formar um conjunto capaz de representar a maior parte dos documentos que serão classificados. Eles devem também elicitar de maneira consistente as características principais de como cada tópico pode ser identificado nos documentos.

A fim de evitar a necessidade de interação humana e possibilitar o treinamento de um modelo de classificação sem um conjunto de treinamento disponível, alguns estudos exploram informações disponíveis na web, como a Wikipédia (YUN et al, 2010) e Wordnet (LUO; CHEN; XIONG, 2011) entre outros (QI; DAVISON, 2009). Nesses trabalhos, os dados foram utilizados para obtenção de conceitos e relacionamentos para uso durante treinamento de classificadores de documentos. Como tais recursos são alimentados continuamente, poderiam se tornar fontes constantes de novos conhecimentos para sistemas de classificação. Contudo, para usufruir dessa característica, a abordagem tradicional precisaria ser complementada com mecanismos capazes de atualizar e adicionar informações ao modelo de classificação, e realizar as etapas de treinamento e validação do classificador constantemente, o que não é desejável.

Inspirada nesses trabalhos, esta dissertação busca esclarecer se é possível resolver o problema de como trabalhar com bases de conhecimento para classificação de documentos, sem a necessidade de emprego de algoritmos de ML e NLP, além de verificar se é possível alcançar resultados próximos aos alcançados pelas técnicas mais comumente empregadas. O emprego de uma abordagem como essa soma os benefícios da utilização de fontes de dados abertos em tempo de classificação com a redução de custo provida pela ausência de treinamento, possibilitando ainda a adição de conhecimento novo de forma independente.

Como o problema geral da classificação de documentos pode ser dividido em problemas mais específicos, tais quais a classificação de sentimento, filtragem de spam, classificação em tópico, entre outros

(AGGARWAL; ZHAI, 2012), torna-se necessário limitar o escopo de atuação para esta dissertação. Desta forma, este trabalho tem foco na classificação em tópico, que visa a descoberta sobre o assunto ou tema de um documento. Por exemplo, julgar se um documento é sobre "futebol", "música" ou "vestuário" é um exemplo de classificação em tópico.

A fim de efetivar a classificação, é assumindo que é possível prever o tópico de um documento através da extração de termos relevantes encontrados nele, e a posterior associação desses termos com um ou mais conceitos de maior nível de abstração, existentes em uma base de conhecimento. A intuição é que os relacionamentos encontrados entre termos do documento e conceitos da base de conhecimento sejam úteis para determinar o assunto de um documento. Por exemplo, se um documento menciona os termos "ator", “roteiro” e "Jack Nicholson", existe uma boa chance de se tratar de um documento sobre filmes.

1.1 OBJETIVO

O objetivo deste trabalho é apresentar uma técnica para classificação de documentos por meio de uma abordagem baseada unicamente em conhecimento.

1.1.1 Objetivos Específicos

A realização do objetivo geral depende do atendimento dos seguintes objetivos específicos:

- Identificar as características desejáveis de uma abordagem para a classificação de documentos em tópicos de forma não supervisionada.

- Propor uma nova abordagem para a classificação de documentos.

- Avaliar a abordagem proposta por meio de experimentos utilizando um conjunto de documentos disponível publicamente.

- Comparar quantitativamente o resultado da classificação com a técnica proposta em relação a métodos tradicionais de classificação de documento.

1.2 CONTRIBUIÇÕES

As contribuições desta dissertação dizem respeito à concepção e avaliação de uma técnica para classificação de documentos em tópicos com a utilização de uma base de conhecimento. A aplicação da técnica proposta oferece vantagens como: (i) ausência da necessidade de emprego de algoritmos de ML ou NLP; (ii) classificação sob demanda; (iii) capacidade de classificar conjuntos de documentos de qualquer tamanho; (iv) oportunidade de atribuir múltiplas classes a um documento; (v) possibilidade de extensão do vocabulário empregado para a classificação sem alteração no classificador; (vi) capacidade de aplicação em diferentes temas de classificação sem a necessidade de adaptação do classificador.

1.3 MÉTODO

Os objetivos são realizados de acordo com as seguintes etapas:

- Estudar os conceitos e as diferentes formas de classificação de documentos em tópicos existentes na literatura.
- Buscar no estado da arte da literatura, abordagens não supervisionadas para classificação de documentos em tópicos e temas afins.
- Identificar oportunidades de inovação e desenvolver um protótipo capaz de realizar de forma automática a classificação de documentos em tópicos.
- Obter um conjunto de documentos reais para avaliar o desempenho do classificador e aprimorá-lo.
- Estudar e desenvolver uma estrutura necessária para a realização de experimentos e coleta e armazenamento dos resultados provenientes desses experimentos.
- Realizar experimentos com o conjunto de documentos obtido utilizando o método proposto e métodos tradicionais de classificação de documentos.
- Analisar os resultados obtidos de acordo com as métricas convencionais existentes na literatura.

- Identificar problemas na abordagem proposta e propor oportunidades de melhoria e trabalhos futuros.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os fundamentos do problema de classificação. O Capítulo 3 apresenta a técnica proposta, detalhando as definições adotadas. Os experimentos realizados para fins de validação são apresentados no Capítulo 4, bem como a discussão dos resultados obtidos. No Capítulo 5 são apresentados os trabalhos mais recentes e relacionados com a abordagem proposta, além de um quadro comparativo com a abordagem apresentada neste trabalho. Conclusões e trabalhos futuros são expostos no Capítulo 6.

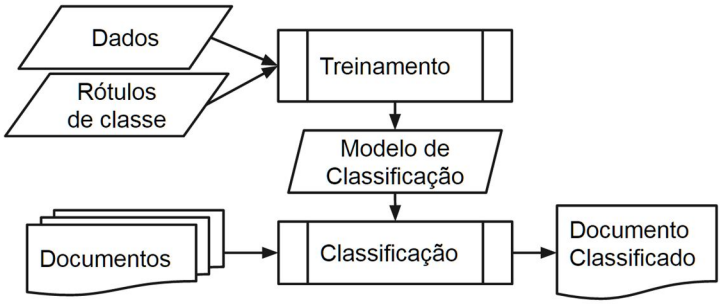
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados conceitos sobre classificação de documentos, as aplicações da tarefa de classificação de documentos, tipos de classificadores e bases de conhecimento.

2.1 CLASSIFICAÇÃO

O problema de classificação de documentos é comumente tratado com abordagens de aprendizagem supervisionada, em que um conjunto predeterminado de dados é usado para treinar um classificador que atribui um ou mais rótulos de classe para cada documento (TAO; LI; LAU; WANG, 2012). Formalmente, o problema é definido como segue (AGGARWAL; ZHAI, 2012). Dado um conjunto de dados para treinamento *D = {d1, ... ,dn}* de modo que cada registo possui um valor de classe extraído de um conjunto de k valores discretos indexados por *{1... k}*. Os dados de treinamento são utilizados para construir um modelo de classificação que relaciona cada registro do conjunto D a um ou mais rótulos de classe. Para uma dada instância de teste em que a classe é desconhecida, o modelo de classificação é utilizado para prever um rótulo de classe.

Figura 1 - Abordagem convencional para classificação de documentos.



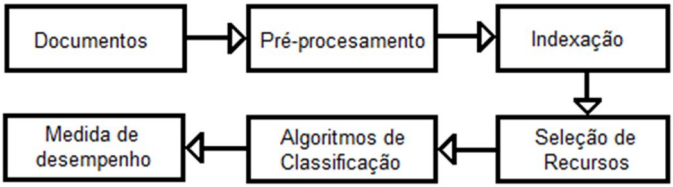
Fonte: Elaborado pelo autor

A **Figura 1** apresenta o problema como é comumente tratado. Em um primeiro momento, é realizada a etapa de treinamento, quando dados de treinamento previamente coletados, escolhidos e preparados são combinados com rótulos de classes a fim de gerar um modelo de classificação. Posteriormente, o modelo de classificação obtido é utilizado para classificar um conjunto de documentos para o qual o modelo de classificação foi definido.

2.1.1 Processo de classificação

O processo moderno de classificação comumente empregado pode ser dividido em etapas como mostra a **Figura 2**.

Figura 2 - Processo convencional de classificação de documentos.



Fonte: Korde (2012)

A primeira etapa diz respeito à coleta de documentos de texto independentemente do formato. A etapa de pré-processamento é responsável por fornecer uma representação dos documentos em formato de texto puro, e geralmente se divide em *tokenization*, limpeza e *stemming*. A etapa de *tokenization* é utilizada para decompor o documento em cada termo que o compõe. A etapa de limpeza diz respeito à remoção de palavras insignificantes e que ocorrem frequentemente como, "as" e "os" por exemplo. *Stemming* é um algoritmo para redução de um termo ao seu radical, removendo as desinências, afixos e vogais temáticas.

O próximo passo no processo de classificação é a indexação dos documentos. Nessa etapa, é necessária a utilização de técnicas que reduzam a complexidade dos documentos e os tornem mais fáceis de manipular. O formato de representação de documentos comumente usado é chamado de modelo de espaço de vetores (AAS; EIKVIL, 1999), onde documentos são representados por vetores de palavras. Esse modelo pode apresentar problemas como perda de correlação com palavras adjacentes e perda de relacionamentos semânticos existentes entre termos do documento.

Para amenizar o problema, geralmente são associados pesos aos termos que formam os vetores. Existem várias formas de associar pesos aos termos como ponderação booleana, peso por frequência, TF-IDF (SALTON; GERARD; MCGILL, 1983), etc. Entretanto a grande desvantagem deste modelo é que ele resulta em uma enorme matriz, o que remete ao problema da alta dimensionalidade. Outros métodos são apresentados por Harish (2010) como, por exemplo: 1) representação de documento em ontologias para manter a relação semântica entre seus termos; 2) uma sequência de símbolos (byte,

caractere ou palavra) chamado de *N-Grams*, que são extraídos de uma longa sequência de um documento; 3) *Latent Semantic Indexing* (LSI), que preserva as características mais representativas em vez de discriminar características.

Mais recentemente, soluções alternativas à representação do conteúdo de documentos por palavras-chave vêm sendo apresentadas, com vistas à adição de semântica e redução do custo computacional. Em um desses trabalhos, Albitar (2012) evidencia que ignorar a semântica de documentos influencia negativamente na classificação. Já Barla (2013) e Poria (2014) concluem que a representação utilizando conceitos-chave é mais eficiente que as técnicas convencionais de representação de conteúdo de documentos.

Após a indexação são aplicadas técnicas de seleção de atributos *(feature selection*) para aprimorar a escalabilidade, eficiência e precisão do classificador. A ideia é manter apenas as palavras com maior escore de acordo com uma medida de importância da palavra, selecionando os recursos que melhor representam o documento para classificação. Vários algoritmos já foram propostos, alguns dos quais são apresentados por Dasgupta (2007) e Zhao (2010).

A etapa de classificação de documentos pode ser realizada de três formas distintas: a forma supervisionada, não supervisionada e semi-supervisionada. Diversas abordagens são implementadas com aplicações de técnicas de ML como o classificador Bayesiano, Árvores de decisão e KNN. Essas e outras técnicas são descritas em maiores detalhes por Korde (2012) e Aggarwal (2012).

Por fim, a classificação é avaliada quanto a sua precisão. Precisão é a medida dada pela capacidade do classificador em associar documentos a classes de acordo com o esperado. Entretanto, é possível que o classificador seja incapaz de atribuir alguma classe para um ou mais documentos. Sendo assim, é necessário medir a capacidade do classificador em termos de quantidade de documentos que ele deveria ser capaz de classificar. A essa medida é dada o nome de cobertura (LEWIS, 1992). Para a obtenção de estimativas de precisão e cobgleidsonertura com relação ao conjunto inteiro de categorias existe a técnica de Medida-F, entre outras.

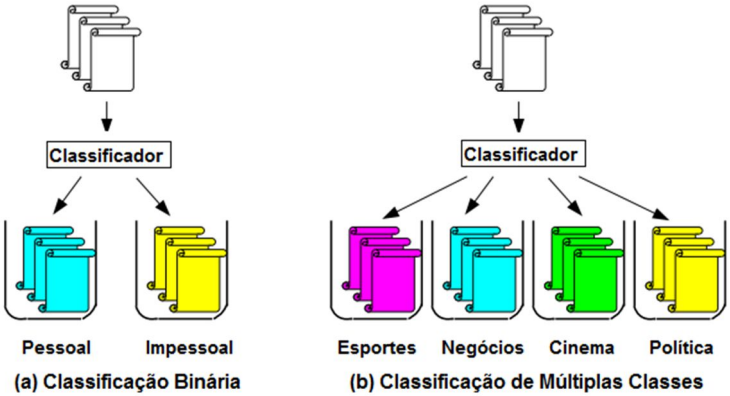
2.1.2 Tipos de classificadores

Na versão tradicional do problema, um rótulo de classe específico é explicitamente atribuído a uma única instância. Entretanto, existem ainda outras variações que ampliam as possibilidades de utilização de classificação. Segundo Xiaoguang (2009), essas variações são consideradas subproblemas do

problema geral de classificação, como por exemplo, a classificação em tópicos, classificação funcional e classificação de sentimentos.

A classificação em tópicos diz respeito ao assunto ou tópico ao qual o documento se refere. Por exemplo, quando se deseja julgar se o conteúdo de um documento é sobre arte, negócio ou esporte. A classificação funcional preocupa-se em identificar o papel que um documento desempenha, por exemplo, determinar se um documento é uma página web pessoal ou uma página web de uma organização. Já a classificação de sentimento foca em identificar a opinião que é apresentada no documento, como descobrir o posicionamento de um autor perante um determinado assunto.

Figura 3 - Classificação binária e de múltiplas classes.



Fonte: Xiaoguang (2009)

Quanto ao número de classes envolvidas no problema, a classificação pode ser dividida em classificação binária e classificação de múltiplas classes. A classificação binária categoriza instâncias em exatamente duas classes (como na **Figura 3** (a)). Já a classificação de múltiplas classes trabalha com mais de duas classes (**Figura 3** (b)). Em relação ao número de classes (rótulos) que podem ser associadas a uma instância, a classificação pode ser dividida em classificação de único rótulo ou de múltiplos rótulos. Único rótulo significa que uma instância pode estar associada exclusivamente a uma classe. Já a classificação de múltiplos rótulos permite que mais de uma classe seja associada a uma única instância.

Com base na forma de atribuição de classe, a classificação pode ser subdividida em classificação rígida e classificação flexível. Na classificação

rígida, uma instância está ou não está em uma classe, sem qualquer estado intermediário, enquanto na classificação flexível, uma instância pode ser atribuída para alguma classe com alguma probabilidade (muitas vezes, uma distribuição de probabilidade entre todas as classes).

Quanto à organização de categorias, a classificação pode ser dividida em plana e classificação hierárquica. Na classificação plana, categorias são consideradas em paralelo, ou seja, uma categoria não substitui a outra. Já na classificação hierárquica, as categorias são organizadas em uma estrutura de árvore hierárquica, em que cada categoria pode ter um número de subcategorias.

2.1.3 Aplicações

Existem diversas aplicações que são de alguma forma beneficiadas pelo uso de classificadores. Alguns exemplos de domínios de aplicação em que a classificação de documentos é amplamente usada incluem (AGGARWAL; ZHAI, 2012, XIAOGUANG; DAVISON, 2009):

- Filtragem e organização de notícias: A maioria dos serviços de notícias hoje é de natureza eletrônica, e um grande volume de artigos de notícias é criado todos os dias pelas organizações. Em tais casos, é difícil organizar as notícias manualmente. Assim, os métodos automatizados podem ser muito úteis para a categorização de notícias em uma variedade de portais web (LANG, 1995). Esta aplicação é também referida como filtragem texto;

- Organização e Recuperação de Documentos: Uma variedade de métodos pode ser usada para a organização de documentos em diversos domínios. Desta forma, esta aplicação inclui grandes bibliotecas digitais de documentos, páginas web, literatura científica, sistemas corporativos, ou até mesmo *feeds* sociais. A organização hierárquica de coleções de documentos pode ser particularmente útil para aprimorar a navegação e recuperação sobre elas (CHAKRABARTI; AGRAWAL; RAGHAVAN, 1997);

- Mineração de Opinião e Análise de Sentimentos: Comentários ou opiniões muitas vezes podem ser considerados como documentos de texto curto, que podem ser extraídos para

determinar informações úteis após serem classificados e analisados (PAK; PAROUBEK, 2010);

- Classificação de E-mail e Filtragem de Spam: Muitas vezes, é desejável classificar e-mails de forma automatizada (CLARK; KOPRINSKA; POON, 2003), a fim de determinar o assunto ou para determinar se o e-mail é indesejado (SAHAMI; DUMAIS; HECKERMAN; HORVITZ, 1998).

2.2 BASES DE CONHECIMENTO

Tradicionalmente, bases de conhecimento são desenvolvidas para prover meios flexíveis a fim de criar, examinar e modificar coleções de estruturas simbólicas (BRODIE; JOHN, 2012). Bases de conhecimento são comumente encontradas como partes que ajudam a compor os chamados sistemas especialistas. Essa organização oferece algumas vantagens: (i) a base de conhecimento é mais fácil de ser lida e atualizada; (ii) os mecanismos de inferência podem ser mais facilmente descritos. Bases de conhecimento "especialistas" são consideradas frágeis. Além disso, o desempenho dessas bases normalmente decai quando utilizadas para resolver outros problemas para os quais não foram projetadas (RECHENMANN, 1995).

Nos últimos anos, grandes bases de conhecimento que não restringem domínio vêm sendo formadas e disponibilizadas para o uso público. Alguns exemplos incluem a DBpedia (LEHMANN, J. et al, 2014), Freebase (BOLLACKER et al, 2008) e YAGO2 (HOFFART, J. et al, 2013). Tais tecnologias se beneficiam direta ou indiretamente do conhecimento gerado pela comunidade Web, tanto para alimentação de dados quanto para melhoria e avaliação de qualidade das informações. Para isso, permitem a participação de usuários que atuam como curadores das informações e também exploram dados da Wikipedia (https://www.wikipedia.org) para extração de conhecimento.

As características de capacidade de extensão do conhecimento, disponibilidade para uso público, grande volume de informações e representação do conhecimento de diferentes domínios fazem com que essas tecnologias constituam importantes fontes de conhecimento para emprego em diferentes tipos de aplicações. O princípio básico de sistemas que fazem uso de bases de conhecimento é justamente a separação explícita entre conhecimento e o mecanismo que o utiliza.

3 KDC: KNOWLEDGE-BASED DOCUMENT CLASSIFICATION

Este capítulo apresenta uma contextualização inicial para a melhor compreensão da abordagem proposta, chamada de KDC (do inglês *Knowledge-based Document Classification*). Em seguida, é apresentada uma visão geral do KDC, e por fim, é detalhado como as definições e conceitos adotados pelo trabalho são utilizados no problema de classificação automática de documentos.

A técnica utiliza termos extraídos a partir de documentos e associa-os a instâncias de conceitos provenientes de uma base de conhecimento de domínio aberto. As instâncias são então generalizadas para conceitos de maior nível de abstração. Os conceitos generalizados são então considerados como rótulos de classe. Para cada conceito é calculado um valor de disparidade com relação ao documento, que é usado para comparar os conceitos de forma a ranqueá-los. Após o ranqueamento, o conceito com menor valor de disparidade é considerado como classe que pode ser aplicável ao documento.

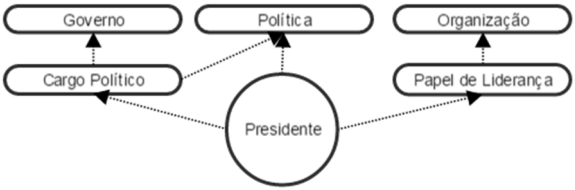
3.1 CONTEXTUALIZAÇÃO

Ao contrário das abordagens de classificação comumente empregadas, o KDC visa proporcionar uma função de classificação sem a dependência de um conjunto de treinamento disponível. Desta forma, não necessita do conhecimento prévio de quaisquer informações sobre os documentos a serem classificados. Esse problema é referido como "classificação não supervisionada de documentos em múltiplas classes". A abordagem proposta utiliza as informações de uma base de conhecimento para realizar tal tarefa. Para explicar claramente a proposta, primeiro se faz necessário introduzir três das principais definições utilizadas: (i) conceito; (ii) base de conhecimento; e (iii) vetor de termos conceituais.

Definição 1. (*Conceito*) Seja C o conjunto de conceitos em uma base de conhecimento e $c \in C$ um par $c := <rótulo, ancestral>$, onde:

- *rótulo* é uma cadeia de caracteres que descreve c;
- *ancestral* refere-se ao conjunto de conceitos diretamente ou indiretamente relacionado a c e localizado em um nível de abstração maior na base de conhecimento, excluindo-se conceitos raiz, tais como “RAIZ”, ou “COISA” por exemplo. Ainda, $ancestral(c) \subset C$.

Figura 4 – Ilustração de um conjunto de conceitos.



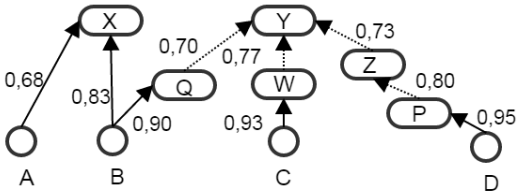
Fonte: Elaborado pelo autor

A **Figura 4** apresenta um exemplo de conjunto de conceitos (elipses). Neste exemplo, os conceitos "Papel de Liderança" e "Cargo Político" não são referenciados por outros conceitos de menor nível de abstração, podendo ser referenciados somente por instâncias de conceitos (círculo) como "Presidente". Por isso, não são considerados ancestrais de nenhum outro conceito. Diferentemente, conceitos como "Governo", "Política" e "Organização" possuem uma maior capacidade de abstração e podem ser usados para representar conceitos relacionados. De acordo com o exemplo, um ancestral de "Cargo Político" é o conceito "Governo".

Definição 2. (*Base de Conhecimento*) Seja B uma base de conhecimento. B contém um conjunto de conceitos relacionados semanticamente em uma estrutura hierárquica e um conjunto de instâncias de conceitos I onde cada $i \in I$ possui um relacionamento *IS-A* com pelo menos um $c \in C$ e para cada relacionamento existe um valor de medida de relevância. Assim, B pode ser definida como uma sêxtupla $B := \langle I; C; R; H_C^R; W; relevância \rangle$, onde:

- I é o conjunto de instâncias de conceitos;
- C é o conjunto de conceitos definido na Definição 1;
- R é o conjunto de relacionamentos entre pares de conceitos;
- H_C^R é a estrutura hierárquica de B construídos por $C \times R$;
- W é o conjunto de relacionamentos entre instâncias e conceitos;
- *relevância* refere-se ao conjunto de valores que denota a força de ligação de cada relacionamento em W, $relevância(i,c) \subset \mathbb{R}$.

Figura 5 – Ilustração de uma base de conhecimento.



Fonte: Elaborado pelo autor

A **Figura 5** ilustra uma base de conhecimento. Nela, círculos representam instâncias de conceitos, elipses representam conceitos e, para cada instância, existe um valor de relevância referente aos conceitos aos quais a instância é diretamente ou indiretamente relacionada. De acordo com a figura, a instância C refere-se aos conceitos W e Y, com valores de relevância de 0.93 e 0.77, respectivamente.

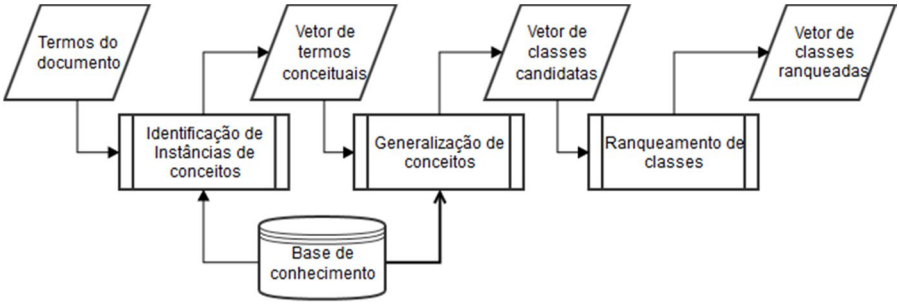
Definição 3. (*Vetor de Termos Conceituais*) Seja D um documento e D_{vetor} um vetor de termos obtido a partir de D onde $D_{vetor} \neq \emptyset$. É definido um vetor de termos conceituais T_{vetor} como um subconjunto de instâncias de uma base de conhecimento $I=\{i_1, ..., i_x\}$ onde, para cada $t \in T_{vetor}$, existe um rótulo único que corresponda ao mesmo tempo a um termo presente em D_{vetor} e uma instância presente em I.

A ideia do vetor de termos conceituais é de usar um conjunto de termos que possa incluir instâncias de conceitos e entidades nomeadas para representar documentos, uma vez que ambas as formas de representação já foram empregadas com sucesso em outros estudos (HUSBY; STEPHANIE; BARBOSA, 2012; BARLA; MICHAL; BIELIKOVA, 2013; LUCIA; FERRARI, 2014).

3.2 VISÃO GERAL DA PROPOSTA

A **Figura 6** mostra as duas etapas principais que são executadas de acordo com a abordagem proposta: a etapa de generalização de conceitos e a de ranqueamento de classes. No primeiro passo, o vetor de termos conceituais é obtido a partir da correspondência entre os termos do documento e instâncias de conceito da base de conhecimento. Em seguida, as instâncias de conceito obtidas são generalizadas. Durante a etapa de generalização de conceitos, cada instância derivada do documento é associada a um conjunto de conceitos de um nível maior de abstração através de consulta à base de conhecimento.

Figura 6 – Processo de classificação proposto.



Fonte: Elaborado pelo autor

A intenção é de fornecer um vetor contendo conceitos mais amplos que podem estar relacionados com o documento. Este vetor é referido como vetor de classes candidatas. Cada classe candidata tem um valor de relevância em relação à instância generalizada. Na segunda etapa, cada classe candidata recebe um valor de disparidade calculado com o uso dos valores de frequência e de distância. Em seguida, as classes candidatas são ranqueadas de acordo com a disparidade. Finalmente, a classe melhor classificada é considerada como o rótulo classe para o documento.

Um exemplo hipotético com vetores de classes candidatas para três termos conceituais é ilustrado na **Figura 7**. Neste exemplo, Presidente, Estado e País são termos obtidos a partir de um documento, para os quais existem instâncias equivalentes presentes na base de conhecimento. Os termos entre parênteses representam o vetor de classes de candidatos para cada termo.

Figura 7 – Exemplo de vetor de termos conceituais e seus respectivos vetores de classes candidatas.

Presidente: {Música, Política, Governo, Organização, Livro}
Estado: { Localização, Organização, Governo, Música}
País: {Governo, Organização, Localização, Música }

Fonte: Elaborado pelo autor

3.3 DEFINIÇÃO DE CLASSES CANDIDATAS

Classes candidatas são definidas utilizando uma função de generalização que mapeia cada instância de um vetor de termos conceituais a um ou mais conceitos presentes na base de conhecimento.

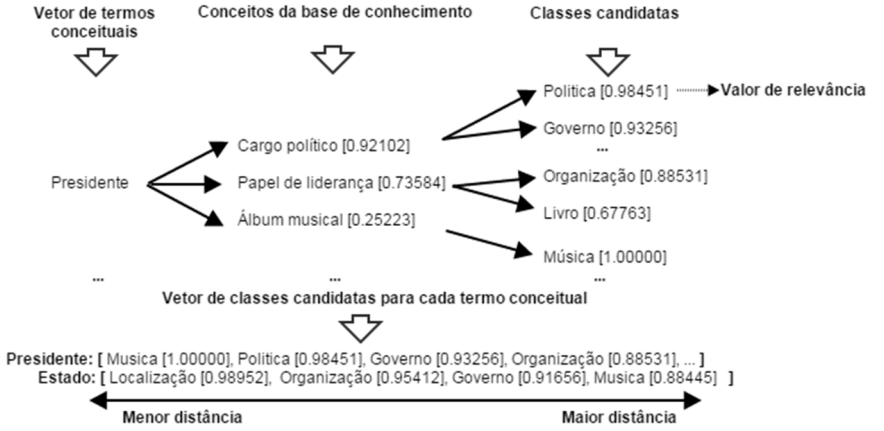
Definição 4. (*Generalização de Conceitos*) Dado um vetor de termos conceituais $T_{vetor}=[t_1,...,t_j]$ obtido de um documento D e uma base de conhecimento B onde, para cada $t \in T_{vetor}$, existe um conjunto de conceitos relacionados $K=\{k_1, ..., k_z\}$ onde $K \subset C$, uma função de generalização de conceitos é definida como *generalize*(B, t)-> α onde α corresponde à união de todos os conjuntos *ancestral* de cada $k \in K$.

$$\alpha = \bigcup_{z=1}^{n} [k_z] \quad (1)$$

A função de generalização usa a estrutura hierárquica da base de conhecimento para obter conceitos de maior nível de abstração relacionados a cada termo conceitual do documento. Para fazer isso, conceitos mencionados direta ou indiretamente por instâncias são consultados. Depois de obter os conceitos mais abstratos, alguns conceitos são selecionados como classes candidatas.

Definição 5. (*Vetor de Classes Candidatas*) Dado um vetor de termos conceituais $T_{vetor}=[t_1, ..., t_j]$ obtido a partir de um documento D e uma base de conhecimento B, onde para cada $t \in T_{vetor}$ existe um conjunto de conceitos *ancestral* $\alpha = \{k_1, ..., k_d\}$, um vetor de classes candidatas $\mathbb{C} = [c_1, ..., c_v]$ é definido como um subconjunto de α contendo todos os conceitos de α que têm um conjunto *ancestral* vazio (*ancestral*(k) = Ø), ordenado por *relevância*(t, k).

Figura 8 – Exemplo de generalização de conceito e derivação de classes candidatas.



Fonte: Elaborado pelo autor

Como exemplo, a estrutura hierárquica mostrada na **Figura 8** apresenta alguns conceitos obtidos a partir de uma base de conhecimento relacionados com o termo conceitual Presidente, incluindo Política, Álbum musical e Governo, além de outros conceitos mais abstratos. A partir dos conjuntos de conceitos generalizados, os vetores de classes candidatas são gerados para cada instância do vetor de termos conceituais.

Para cada termo conceitual, um vetor de classes candidatas é criado. Para fazer isso, os conceitos de maior nível de abstração, excluindo-se os conceitos raízes, são selecionados. Estes conceitos são agrupados para formar um vetor ordenado por valor de relevância obtido da base de conhecimento. Para ilustrar este processo, a **Figura 8** mostra os conceitos generalizados e a seleção de classes candidatas para a instância Presidente. Entre os conceitos generalizados, aqueles que não têm nenhum conceito *ancestral*, incluindo Música, Politica, Governo e Organização são usados para formar um vetor ordenado pelo valor de relevância.

3.4 RANQUEAMENTO DE CLASSES CANDIDATAS

Após definir o vetor de classes candidatas, o KDC efetua o cálculo de uma medida de disparidade utilizando os valores de distância e frequência de cada classe candidata. Nesta proposta, a frequência de uma classe candidata equivale ao número de vezes em que ela é referenciada, considerando-se todos os vetores de classes candidatas de todos os termos conceituais do documento. A intuição é que a frequência fornece um valor que expressa o quão importante cada classe pode ser no contexto do documento.

Definição 6. (*Frequência*) Seja $\mathcal{C}_{set} = \{\mathcal{C}_1, ..., \mathcal{C}_x\}$, o conjunto de todos os vetores de classes candidatas do vetor de termos conceituais, a *frequência*(ϵ, $\mathcal{C}_{set}$) $\in \mathbb{N}$ é definida como uma função que conta quantas vezes uma classe candidata ϵ_i aparece em $\mathcal{C}_{set}$. Quanto maior o valor obtido, mais importante para o documento o candidato é.

Deste modo, de acordo com o exemplo da **Figura 9**, a frequência da classe Localização seria 2, porque ela está contida em dois vetores de classes candidatas.

Figura 9 – Exemplo de vetores de classes candidatas.

Instância	Vetor de classes candidatas				
Presidente	Música	Política	Governo	Organização	Livro
Estado	Localização	Organização	Governo	Música	
País	Governo	Organização	Localização	Música	
Índice	**1**	**2**	**3**	**4**	**5**

Fonte: Elaborado pelo autor

Considera-se distância, a soma das posições de cada classe candidata nos vetores de classes candidatas. O valor de *distância* denota a probabilidade de uma classe ser mais importante que as demais.

Definição 7. (*Distância*) Seja ϵ uma classe candidata, $V_{class}=\{v_1, \ldots, v_n\}$ um conjunto de n vetores de classes candidatas contendo ϵ, onde n corresponde ao valor de frequência, e $VP = [pcv_1, \ldots, pcv_n]$ um vetor onde cada valor corresponde à posição de ϵ em cada v_n. O valor de distância pode ser calculado por:

$$distância(\epsilon) = \sum_{j=1}^{n}[pcv_j - 1] \quad (2)$$

Como exemplo, considerando a classe candidata Localização da **Figura 9**, é observado que, a aplicação da **Equação 2** resulta em um valor igual a 2. Uma vez obtidos, os valores de frequência são comparados entre as classes para dar origem a um vetor ranqueado.

Definição 8. *(Ranque de Frequência)*: Seja $F=\{f_1,...,f_z\}$ um conjunto em que f é definido pelo par $(\epsilon, frequência(\epsilon))$, onde existe um f para todo $\epsilon \in \mathbb{C}_{set}$. É considerado ranque de frequência o valor obtido a partir de uma função $fRank(\epsilon, F) \in \mathbb{N}$ que compara ϵ com todos os elementos presentes em F de acordo com o valor de $frequência(\epsilon)$, onde os itens que sejam considerados iguais ocupam a mesmo posição no ranque.

A posição de cada classe no ranque de frequência pode ser usada para denotar o quão próximo ao documento a classe é. De acordo com as definições estabelecidas, as classes candidatas com maior valor de frequência terão seu *fRank* igual a 1, enquanto classes com frequências menores terão valores maiores. Para ilustrar, considerado o conceito Localização, de acordo com a **Figura 9**, como seu valor de frequência é 2 e esse é o segundo maior valor de frequência entre as classes candidatas, seu ranque de frequência é 2.

Após encontrar os valores de frequência e distância, um valor de disparidade pode ser calculado. O valor de disparidade de uma classe candidata considera a soma do ranque de frequência e o valor de distância. O resultado, dividido pela frequência, estabelece uma distância média com a influência da frequência.

Definição 9. (*Disparidade*): Seja $\mathfrak{c}$ uma classe candidata, $\mathcal{C}_{\text{set}} = \{\mathcal{C}, ..., \mathcal{C}_{\text{x}}\}$ o conjunto de todos os vetores de classes candidatas de um vetor de termos conceituais, e $F = \{f_1, ..., f_{\text{z}}\}$ um conjunto onde f é definido pelo par ($\mathfrak{c}$, frequência($\mathfrak{c}$)) onde existe um f para todo $\mathfrak{c} \in \mathcal{C}_{\text{set}}$. Considera-se disparidade o valor dado por:

$$disparidade(\mathfrak{c}, \mathrm{F}, \mathrm{Z}) = \frac{(distância(\mathfrak{c}) + fRank(\mathfrak{c}, \mathrm{F}) - 1)}{frequência(\mathfrak{c}, \mathcal{C}\mathrm{set})} \qquad (3)$$

Tabela 1**.** Exemplos de valores calculados.

Classe candidata	Frequência	F- Rank	Distância	Disparidade
Organização	3	1	5	1.66
Governo	3	1	4	1.33
Música	3	1	6	2
Localização	2	2	2	1.5
Livro	1	3	4	6
Política	1	3	1	3

Fonte: Elaborado pelo autor

Essa equação resulta em valores próximos de zero quando a classe possui maior afinidade com o documento. Para exemplificar, considere os valores obtidos a partir dos exemplos anteriores com relação à distância, frequência e ranque de frequência. O valor de disparidade para a classe candidata Localização seria dada por 2+2-1/2=1,5. Os valores de disparidade para todos os exemplos de classes presentes na **Figura 9** estão presentes na **Tabela 1**.

Depois de calcular os valores de disparidade para todas as classes candidatas, a classe com o menor valor de disparidade é considerada como rótulo do documento. Se duas ou mais classes possuem o menor valor do vetor,

então os valores de frequência e distância são usados respectivamente como critérios de desempate.

4 AVALIAÇÃO EXPERIMENTAL

Neste capítulo, são descritos os experimentos realizados para demonstrar a efetividade da proposta deste trabalho utilizando documentos reais para a realização de classificação.

4.1 CONFIGURAÇÃO DO CONJUNTO DE DADOS

Os experimentos foram realizados com o uso de um subconjunto de documentos do DMOZ (*The Open Directory Project*, http://www.dmoz.org). Nesses experimentos, o Freebase (https://www.freebase.com) foi utilizado como base de conhecimento. A extração de termos a partir dos documentos foi realizada com a ferramenta de mercado AlchemyAPI (http://www.alchemyapi.com). A associação das categorias do DMOZ com os domínios do Freebase foi realizada como apresentado na **Tabela 2**.

Tabela 2**.** Categorias do DMOZ e os domínios do Freebase correspondentes utilizados nos experimentos.

DMOZ	Freebase
Top/Sports/Soccer/	Soccer
Top/Games/Video_Games/Action	Video Games
Top/Arts/Music/Bands_and_Artists/	Music
Top/Shopping/Clothing/	Fashion, Clothing
Top/Health/Medicine	Medicine
Top/Computers/Hardware	Computers
Top/Recreation/Food/Drink	Food & Drink
Top/Arts/Movies/Titles	Film
Top/Recreation/Travel	Location
Top/Reference/Education/Colleges_and_Universities	Education

Fonte: Elaborado pelo autor

Por ser o maior diretório web editado por humanos e prover *dumps* regulares com as informações que possui, o DMOZ foi utilizado como fonte de documentos previamente classificados para a avaliação do KDC. Após a associação das categorias com os domínios do Freebase, foram selecionados documentos que continham no mínimo 50% de termos extraídos pela ferramenta relacionados de alguma forma com o domínio do Freebase desejado. Esse critério foi adotado para assegurar um mínimo de qualidade com relação ao conteúdo das páginas, pois muitas delas eram formadas apenas por imagens, ou apresentavam, por exemplo, maior volume de texto sobre outros assuntos como propaganda. Apesar da existência de técnicas para tratamento deste tipo de conteúdo, a solução aplicada permite que o problema seja resolvido de forma simples, sem a necessidade de desviar o foco dos experimentos. Desta forma, para os experimentos, foram utilizados 9329 documentos distribuídos nas dez diferentes categorias.

Tabela 3**.** Números de conceitos e instâncias presentes nos documentos.

Freebase	Documentos	Conceitos	Instâncias	Instâncias por Conceito
Soccer	490	24	367833	15326
Video Games	1100	32	107517	3360
Music	1100	69	38111746	552344
Fashion, Clothing	604	10	2377	238
Medicine	1100	70	1040479	14864
Computers	689	31	39529	1275
Food & Drink	1100	68	309319	4549
Film	1099	54	5482760	101533
Location	1100	176	5475002	31108
Education	947	30	1079233	35974

Fonte: Elaborado pelo autor

As categorias foram escolhidas por serem distintas entre si, além de possuirem um grande número de documentos e por serem compatíveis com

conceitos de maior nível de abstração do Freebase, que são referenciados como domínio. O Freebase é uma base de conhecimento de domínio aberto com esquemas bem definidos, possui uma função de relevância, mais de 218 milhões de instâncias, 2.9 bilhões de relacionamentos e uma API para utilização de seus recursos. A **Tabela 3** provê informação sobre o número de instâncias e conceitos em cada domínio do Freebase e o número de documentos do DMOZ usados para a avaliação.

4.2 METODOLOGIA

Para alcançar os objetivos propostos, a abordagem foi comparada com três métodos clássicos de classificação de documentos, consideradas como *baselines*: (i) Árvore de Decisão (J48), por obter bons resultados com vetores de pequenas dimensões; (ii) Support Vector Machine (SVM), por obter bons resultados em diversos estudos de classificação de documentos, incluindo dados de alta dimensionalidade; e (iii) Naive Bayes (NB), por apresentar excelentes resultados com o conjunto de dados estudado. Todos os experimentos foram realizados usando as métricas de precisão, cobertura e Medida-F (WITTEN; FRANK; HALL, 2011), que são as métricas comumente usadas para comparar sistemas de classificação de documentos onde:

- Precisão: Consiste em uma relação do número de documentos classificados corretamente em uma classe (Verdadeiros Positivos (Vp)) pelo número total de documentos classificados como pertencentes a essa classe (Verdadeiros Positivos + Falsos Positivos (Fp)).

$$Precisão = \frac{Vp}{Vp + Fp}$$

- Cobertura: Calculada pela divisão do número de Verdadeiros Positivos pelo número real de documentos dessa classe (Verdadeiros Positivos + Falsos Negativos (Fn)).

$$Cobertura = \frac{Vp}{Vp + Fn}$$

- Medida-F (F): É uma média dos valores de Precisão e Cobertura, usada para comparar classificadores por meio de uma única medida.

$$F = \frac{2 * Precisão * Cobertura}{Precisão + Cobertura}$$

O ambiente de avaliação experimental foi configurado da forma como segue nos próximos capítulos.

4.2.1 Configuração do KDC

A fim de aplicar a abordagem proposta, um protótipo que implementa o KDC foi desenvolvido. O protótipo usa a ferramenta AlchemyAPI para extrair termos a partir de documentos e a base de conhecimento Freebase. A escolha de ferramenta de extração de instância foi influenciada pelo bom desempenho apresentado em (GAGNON; MICHEL; ZOUAQ; JEAN-LOUIS, 2013) juntamente com o fato de ter uma versão gratuita para uso acadêmico e disponibilidade de uma API Java. Antes de realizar a classificação, foi necessário configurar o protótipo para considerar apenas as classes utilizadas na experiência.

4.2.2 Configuração de baselines

Depois de classificar documentos usando o protótipo, os arquivos de *log* resultantes foram usados para gerar arquivos em formato ARFF para serem usados com a ferramenta WEKA (HALL et. al, 2009), contendo o mesmo conjunto de instâncias de conceitos utilizado no protótipo. O uso de instâncias de conceitos para representar o conteúdo do documento é considerado válido, uma vez que já foi bem sucedido em outros estudos (HUSBY; STEPHANIE; BARBOSA, 2012; BARLA; MICHAL; BIELIKOVA, 2013; LUCIA; FERRARI, 2014). Para representação do documento, o modelo de espaço vetorial foi utilizado em conjunto com TF-IDF. Para a seleção de atributos (*feature selection*), foram utilizados *Information Gain* e *Chi Square*, pois são os métodos mais usados e que obtém os melhores resultados (KORDE; MAHENDER, 2012).

4.3 RESULTADOS

Para uma melhor apresentação dos resultados obtidos, esta seção apresenta os resultados experimentais divididos em três categorias:

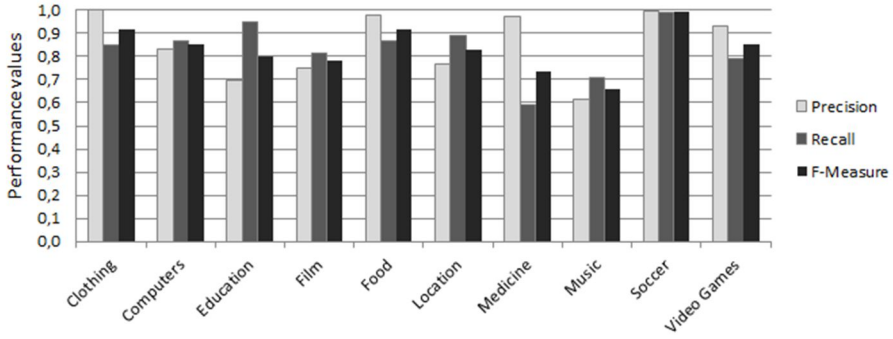
1. Resultados do KDC: descrevem comparativamente os diferentes conjuntos de dados usando a nossa proposta;

2. Resultados de *baselines:* apresentam resultados importantes usando os algoritmos Árvore de Decisão, SVM e Naive Bayes;
3. Resultados comparativos, descrevem a proposta deste trabalho com realação aos *baselines*.

4.3.1 Resultados do KDC

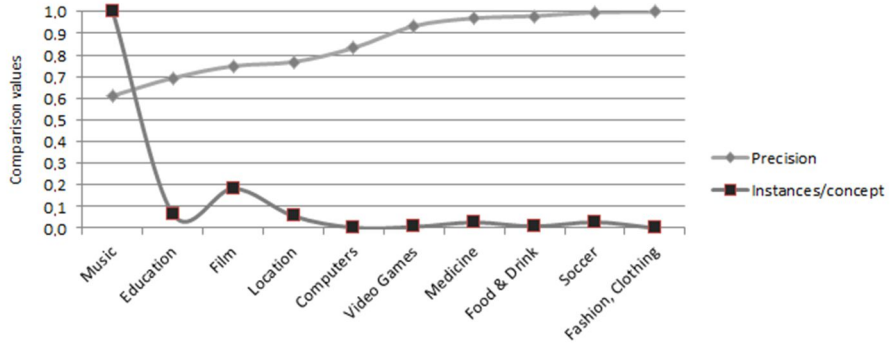
Os resultados de classificação obtidos a partir do protótipo de acordo com as medidas de precisão, cobertura e medida-f são apresentados na **Tabela 4** e na **Figura 10**.

Figura 10 – Gráfico de desempenho do KDC por classe.



Fonte: Elaborado pelo autor

Figura 11– Comparação entre precisão e a proporção de instâncias por conceito.



Fonte: Elaborado pelo autor

A partir dos experimentos realizados pode-se dizer que a classificação utilizando KDC alcança melhores resultados em termos de precisão quando a proporção entre o número de instâncias por conceitos na base de conhecimento

é menor, como pode ser visto na **Figura 11**. Como mostrado na **Tabela 3**, as classes *Medicine*, *Video Games*, *Soccer*, *Food*, *Clothing* e *Computers* possuem valores mais baixos no número de instâncias por conceitos.

As classes mais abrangentes, como *Music*, *Education*, *Film* e *Location* tiveram menor precisão do que o esperado, uma vez que podem envolver um número muito maior de instâncias por conceito na base de conhecimento, o que aumenta as possibilidades de classificação. No entanto, isso é parcialmente compensado pelo fator de cobertura. Uma possível solução para este problema envolve a aplicação do valor de relevância para a seleção dos principais conceitos que podem representar melhor uma instância.

Tabela 4**.** Medidas de desempenho dos experimentos com o KDC

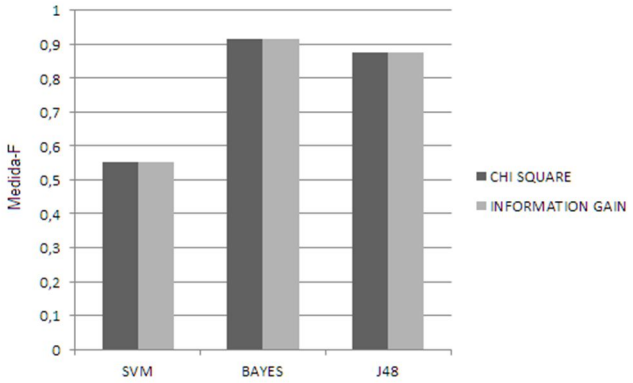
Freebase	Documentos	Precisão	Cobertura	Medida-F
Soccer	490	0.997	0.989	0.993
Video Games	1100	0.934	0.788	0.855
Music	1100	0.612	0.707	0.656
Fashion, Clothing	604	1.000	0.847	0.917
Medicine	1100	0.970	0.593	0.736
Computers	689	0.833	0.869	0.850
Food & Drink	1100	0.980	0.866	0.919
Film	1099	0.749	0.814	0.780
Location	1100	0.768	0.892	0.826
Education	947	0.694	0.951	0.802
Média	**933**	**0.854**	**0.832**	**0.833**

Fonte: Elaborado pelo autor

4.3.2 Resultados dos baselines

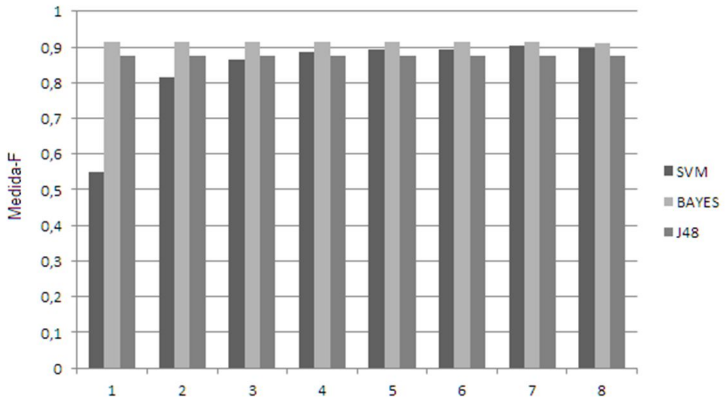
A fim de encontrar os melhores resultados, foram empregadas técnicas de seleção de atributos para cada método de classificação, obtendo resultados muito semelhantes como pode ser visto na **Figura 12**.

Figura 12 – Medida-F dos diferentes métodos usando *Chi Square* e *Information Gain*.



Fonte: Elaborado pelo autor

Figura 13 – Medida-F de acordo com o valor mínimo de frequência para TF-IDF.



Fonte: Elaborado pelo autor

Antes de obter os resultados finais, algumas experiências foram realizadas usando cada método para encontrar o melhor valor mínimo de frequência para o emprego de TF-IDF como mostrado na **Figura 13**. Os resultados mostram que o classificador SVM foi o maior beneficiado pela utilização de um número menor de atributos.

Além disso, cada classificador foi modelado de acordo com o método de divisão percentual com a percentagem variando entre 60% a 90% dos dados sendo usados para treinamento e os restantes 40% a 10% para testar o

classificador. Os resultados obtidos a partir de cada classificador são mostrados na **Tabela 5**, onde J48 e SVM obtiveram melhores desempenhos utilizando a razão de 90/10, enquanto NB funcionou melhor utilizando a proporção 70/30.

Tabela 5**.** Resultados obtidos a partir de métodos de aprendizagem de máquina.

	Precisão			**Cobertura**			**Medida-F**		
	J48	**SVM**	**NB**	**J48**	**SVM**	**NB**	**J48**	**SVM**	**NB**
90%/10%	0.897	0.931	0.915	0.880	0.901	0.908	0.884	0.907	0.909
80%/20%	0.874	0.926	0.918	0.859	0.887	0.910	0.863	0.894	0.912
70%/30%	0.888	0.925	0.919	0.872	0.892	0.911	0.876	0.898	0.913
60%/40%	0.877	0.914	0.916	0.863	0.882	0.908	0.866	0.886	0.910
Média	**0.884**	**0.924**	**0.917**	**0.868**	**0.89**	**0.909**	**0.872**	**0.896**	**0.911**

Fonte: Elaborado pelo autor

A necessidade de tais experimentos, para encontrar os melhores resultados possíveis utilizando aprendizagem de máquina, mostra que não é fácil calibrar um classificador para um conjunto de tamanho limitado de documentos conhecidos. Também é observado que o número de documentos utilizados para treinamento pode impactar diretamente na qualidade dos resultados. Se o tamanho do conjunto de documentos a ser classificado for muito maior que o tamanho do conjunto de treinamento, a utilização de ML pode ser inviável. Outro ponto válido a ser mencionado, é que experimentos preliminares tiveram de ser realizados para que fosse possível definir um número mínimo de documentos a serem usados nos experimentos. Isso foi necessário para que fosse possível produzir melhores resultados com as abordagens de *baseline*, já que o KDC consegue classificar conjuntos de documentos de qualquer tamanho maior que zero.

4.3.3 Comparação de resultados

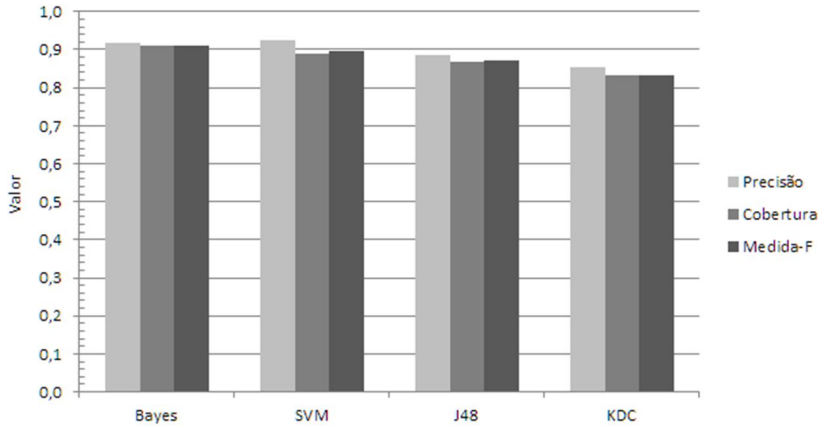
Os melhores resultados obtidos com os métodos clássicos são comparados com os resultados da abordagem proposta na **Tabela 6**.

Tabela 6**.** Comparação entre resultados.

	Precisão	**Cobertura**	**Medida-F**
NB	0.919	0.911	0.913
SVM	0.931	0.901	0.907
J48	0.897	0.880	0.884
KDC	0.854	0.832	0.833

Fonte: Elaborado pelo autor

Figura 14 - Comparação entre as médias das abordagens de aprendizagem de máquina e KDC.



Fonte: Elaborado pelo autor

Embora tenha obtido resultados mais baixos do que os métodos baseados em ML, como pode ser visto na **Figura 14**, a abordagem apresentada neste trabalho pode ser capaz de executar a tarefa de classificação em documentos reais usando uma base de conhecimento. Isso permite que documentos sejam classificados em categorias pré-definidas, sem qualquer informação prévia.

Outros benefícios do KDC incluem a classificação de documentos em tempo real, a possibilidade de enriquecimento do conhecimento utilizado pelo classificador através da inserção de novos conceitos e instâncias na base de conhecimento, e a ausência da necessidade de manter um modelo de

classificação atualizado constantemente, reduzindo a complexidade e a necessidade de manutenção do classificador.

4.3.4 Relevância estatística

A fim de verificar se os métodos avaliados produziram resultados com o mesmo significado estatístico, os valores de Medida-F obtidos foram utilizados em testes estatísticos. Os testes aplicados permitem que mais de dois grupos sejam comparados utilizando variáveis contínuas. Os testes exigem que a variável de interesse tenha distribuição normal e que os grupos sejam independentes (SOUZA, 2015). Desta forma, foram usados três resultados de Medida-F de cada método a partir dos experimentos realizados.

Para os valores de Medida-F do KDC foram usados o menor valor obtido, o valor médio e o maior valor. Para os métodos comparados foram usados os valores de Medida-F obtidos nos testes de variação de instâncias de treino e teste entre 70%, 80% e 90%. Para testar a normalidade dos dados foi utilizado o método Shapiro-Wilk (SHAPIRO; WILK, 1965), utilizado quando existe apenas uma variável a ser analisada. A **Tabela 7** apresenta os resultados obtidos, nela é possível observar que os valores calculados de W são maiores que os valores críticos de W, mostrando que a distribuição é normal.

Tabela 7**.** Resultado do teste de normalidade usando Shapiro-Wilk e Medida-F.

	Medida-F	**W**	**Crit W**
NB	0.909, 0.915, 0.912	1.000000	0.767
J48	0.875, 0.883, 0.865	0.995902	0.767
SVM	0.863, 0.884, 0.878	0.942308	0.767
KDC	0.833, 0.993, 0.656	0.999152	0.767

Fonte: Elaborado pelo autor

Após a confirmação de que o conjunto possui uma distribuição normal, foi possível realizar o teste de análise de variância ANOVA. Esse teste é usado para a comparação de mais de dois conjuntos de valores. Desta forma foram usados os valores dos quatro conjuntos de resultados (NB, J48, SVM e KDC). O resultado obtido ($F_{calc} = 0.504 < F_{tab} = 4.066$), de acordo com os critérios

convencionais, indica que não existem diferenças estatisticamente significativas entre as médias.

Outro teste realizado foi o teste T (STUDENT, 1908), precursor do ANOVA, aplicável a apenas dois grupos distintos. Para esse teste, os resultados obtidos pelo KDC foram comparados com os resultados obtidos com NB, que apresentou os melhores resultados sobre os experimentos realizados. De acordo com os critérios convencionais, o resultado ($P = 0.4335$) do teste T também indica que a diferença entre os resultados não é estatisticamente significativa. Desta forma é possível afirmar que, estatisticamente, para os resultados obtidos com os experimentos realizados, os resultados obtidos pelo KDC são equivalentes aos dos métodos baseados em ML.

5 TRABALHOS RELACIONADOS

A tarefa de classificação de documentos em tópicos de forma não supervisionada, e em múltiplas classes, refere-se à associação de rótulos de classe a documentos sem qualquer informação prévia sobre eles. Além disso, não deve haver necessidade de intervenção humana no processo. Para este fim, várias pesquisas já foram apresentadas aplicando ML e evoluindo seu processo de classificação. Estes estudos exploram técnicas, incluindo a automação da etapa de treinamento do classificador, *clustering, o* uso de ontologias e bases de conhecimento. Este capítulo apresenta os principais trabalhos relacionados a essas abordagens, diferenciando características de cada uma no que diz respeito ao KDC.

5.1 AUTOMAÇÃO DO PROCESSO DE TREINAMENTO

Uma das limitações de abordagens que fazem uso de ML é a necessidade de gerar um modelo de classificação utilizando dados de treinamento. Para reduzir o impacto dessa limitação, alguns estudos foram desenvolvidos com foco na automatização de alguma tarefa da etapa de treinamento e construção do modelo de classificação.

A fim de rotular automaticamente documentos usados na etapa de treinamento de um classificador bayesiano, (KO; SEO, 2000) propõem um método que divide os documentos em sentenças. Posteriormente, todas as sentenças são associadas a uma categoria por meio de uma medida de similaridade, obtida pela comparação das sentenças com listas de palavras-chave, onde cada lista pertence no mínimo a uma categoria.

Para realizar classificação de blogs, (HUSBY; STEPHANIE; BARBOSA, 2012) usa páginas da Wikipédia marcadas automaticamente com domínios do Freebase (BOLLACKER; EVANS; PARITOSH; STURGE; TAYLOR, 2008) como fonte de dados para formar o conjunto de dados de treinamento. A definição dos rótulos é realizada com o uso de um *tagger* convencional. Para realizar a classificação, métodos de aprendizagem de máquina foram testados usando WEKA.

Embora as abordagens apresentadas por estes trabalhos não exijam envolvimento humano para realizar a classificação, algoritmos de classificação convencionais são geralmente mais adequados às coleções de documentos estáticos e não são a solução ideal para coleções de tamanhos dinâmicos (SIGOGNE; CONSTANT, 2009).

5.2 CLASSIFICAÇÃO BASEADA EM CLUSTERING

Outra forma de emprego de ML para a classificação de documentos é agrupá-los em *clusters*. Uma das principais características que atribuem vantagem à utilização de *clustering* é que ela possibilita ao classificador o manuseio de coleções de documentos de tamanhos variáveis.

Entre as abordagens que usam *clustering* para a classificação de documentos de forma não supervisionada, (HERNÁNDEZ; RIVERO; RUIZ; CORCHUELO, 2014) cria uma série de padrões de URL que representam as diferentes classes de páginas em um site. Então, outras páginas podem ser classificadas relacionando suas URLs com os padrões encontrados.

De forma similar, (PANG; JIANG, 2013) emprega um algoritmo de clusterização para treinar um modelo de categorização genérico baseado em clusters. Em seguida, usa a regra de decisão KNN para classificar documentos de texto com base no modelo. Para fornecer a classificação em tempo real de documentos, (SIGOGNE; CONSTANT, 2009) usa o modelo de espaço vetorial para a representação de conteúdo e TF-IDF e ponderação de termos para a classificação de documentos por meio de clusterização.

Apesar de fornecerem maneiras eficientes para classificação de documentos, as abordagens baseadas em *clustering* diferem da abordagem presente neste trabalho por tratar classes apenas como conjuntos de documentos, sem a possibilidade de considerar qualquer distinção semântica explícita de tópicos.

5.3 CLASSIFICAÇÃO UTILIZANDO ONTOLOGIA

Ontologias também foram exploradas para a tarefa de classificação não supervisionada. O trabalho apresentado em (TAO; LI; LAU; WANG, 2012) classifica documentos pela correspondência entre termos e tópicos. Os termos são obtidos a partir de um processo de mineração de dados sobre os documentos que devem ser classificados. Os tópicos são extraídos de uma ontologia de domínio aberto. A ontologia é usada para generalizar os termos adquiridos, a fim de obter conceitos que podem ser utilizados como rótulos de classe para a classificação. Apesar de evitar a necessidade de construção de um modelo de classificação, a abordagem apresentada por eles ainda depende do conhecimento prévio dos documentos que serão classificados.

A abordagem apresentada em (LUCIA; FERRARI, 2014) utiliza uma ontologia em RDF, realizando a classificação de documentos de pequenos volumes de texto. Para isso, a técnica utiliza um valor de pontuação obtido sobre os arquivos RDF. A pontuação é baseada na distância entre as instâncias obtidas a partir do documento e conceitos considerados como categorias, que

devem ser previamente estabelecidas. O KDC difere dessa abordagem por poder lidar com um maior número de termos, o que também permite a classificação de documentos de maior dimensão.

Outro trabalho relacionado é apresentado em (ALLAHYARI; KOCHUT; JANIK, 2014). Nele, páginas da Wikipédia foram convertidas para uma ontologia em RDF. Essa ontologia foi então usada para categorizar um corpus de documentos de notícias. O método usa conjuntos de temas definidos previamente pelo usuário e contextos especificados na ontologia para realizar a classificação. A correspondência do documento em classes definidas pelo usuário é realizada de acordo com a semelhança semântica dos grafos conceituais montados a partir de termos de documentos obtidos nos contextos definidos. A estratégia apresentada nesse trabalho requer um especialista para identificar documentos relevantes, e definir as configurações que podem ser utilizadas para a classificação. Por isso, se encaixa melhor em domínios específicos do que no domínio aberto.

5.4 QUADRO COMPARATIVO

O quadro apresentado na **Tabela 8** sintetiza as principais características das abordagens relacionadas com a abordagem proposta nesta dissertação. A coluna "Treinamento Necessário" indica se a abordagem necessita de uma etapa de treinamento para construção de um modelo de classificação. A segunda, "Intervenção Humana", diz respeito à necessidade de interação humana durante o processo de classificação. A capacidade de associar rótulos de classe predefinidos aos documentos classificados é indicada na coluna "Classifica em Tópico", já a coluna "Conjunto Dinâmico", determina se a abordagem é capaz de lidar com conjuntos de documentos de tamanhos variáveis. Por fim, "Múlti-Rótulos" indica se a técnica atribui mais de um rótulo de classe aos documentos e o "Tipo de Documento" informa quais restrições são impostas para permitir a classificação de documentos com a abordagem correspondente.

A partir do comparativo é possível notar que a maior parte dos trabalhos relacionados não aplica treinamento de modelo de classificação, também não necessita de intervenção humana, além de classificar documentos em tópicos, e aplicar somente um rótulo. Quanto à capacidade de conseguir trabalhar com conjuntos de documentos de tamanhos variáveis e documentos de diferentes tamanhos é observável que quase metade dos trabalhos possui restrições.

Tabela 8**.** Comparativo entre as abordagens relacionadas e este trabalho.

Referências	Treinamento Necessário	Intervenção Humana	Classifica em Tópico	Conjunto Dinâmico	Multi Rótulos	Tipo de Documento
(KO; SEO, 2000)	Sim	Sim	Sim	Não	Não	Apenas Grande
(HUSBY; STEPHANIE; BARBOSA, 2012)	Sim	Não	Sim	Sim	Não	Todos
(HERNÁNDEZ; RIVERO; RUIZ; CORCHUELO, 2014)	Não	Não	Não	Sim	Não	Apenas Web
(PANG; JIANG, 2013)	Sim	Não	Sim	Não	Não	Todos
(SIGOGNE; CONSTANT, 2009)	Não	Não	Não	Não	Não	Todos
(TAO; LI; LAU; WANG, 2012)	Não	Não	Sim	Não	Sim	Apenas Grande
(LUCIA; FERRARI, 2014)	Não	Não	Sim	Sim	Sim	Apenas Pequeno
(ALLAHYARI; KOCHUT; JANIK, 2014)	Não	Sim	Sim	Sim	Não	Todos
KDC	Não	Não	Sim	Sim	Não	Todos

Fonte: Elaborado pelo autor

6 CONCLUSÕES E TRABALHOS FUTUROS

Esta dissertação apresenta o KDC, uma abordagem para a classificação não supervisionada de documentos em tópicos. O KDC explora unicamente recursos semânticos para a tarefa de classificação, sendo uma alternativa viável para quando não existe a possibilidade de emprego de aprendizagem de máquina e a necessidade de classificação de documentos em tópicos.

Como proposto neste trabalho, o KDC é baseado em dois processos principais. O primeiro processo recebe como entrada um conjunto de termos conceituais e fornece um vetor de classes candidatas para que, em um segundo momento, seja eleita uma classe como rótulo, capaz de representar o documento analisado.

Em maiores detalhes, o primeiro processo da abordagem conta com o apoio de uma base de conhecimento. Com este recurso, realiza a generalização de instâncias de conceitos obtidas a partir do vetor de termos conceituais obtidos do documento. Após encontrar os conceitos de maior abstração, o KDC os utiliza como rótulos de classe. Para cada classe, um valor de disparidade é calculado com base nos conceitos relacionados ao documento. Este valor é utilizado para ranquear as classes que podem ser aplicáveis ao documento. Desta forma, o KDC é capaz de realizar a classificação de documentos em tempo real, sem qualquer informação prévia sobre os documentos.

Para a avaliação da abordagem proposta, foram realizados experimentos com um protótipo que implementa o KDC. Como fonte de dados de teste, foi utilizado um conjunto de documentos reais obtidos a partir do DMOZ. Para fins de comparação, os documentos utilizados para classificação com o KDC também foram classificados com o uso das principais abordagens baseadas em ML. A avaliação demonstra que os resultados obtidos no que diz respeito à precisão e cobertura, mostraram-se próximos aos de abordagens tradicionais.

A partir dos experimentos, foi possível demonstrar que quando a proporção entre o número de instâncias e o número de conceitos é menor, os resultados de precisão são melhores. Entretanto, quando o número de instâncias é maior, a cobertura aumenta, mas a precisão é penalizada. Em linhas gerais, este comportamento pode ser explicado pela capacidade de distinção de conceitos entre instâncias da base de conhecimento.

Considerando como exemplo uma base de conhecimento onde exista um grande número de instâncias e conceitos, onde cada instância é associada a um número reduzido de conceitos, e os conceitos são relacionados de forma única entre si, é maior a probabilidade de se obter resultados melhores para a classificação com o KDC. Por outro lado, uma base de conhecimento que contenha um pequeno número de instâncias relacionadas a um grande número de conceitos, que possuam um grande número de relacionamentos entre si,

fornecerá o pior resultado para classificação, pois aumenta as possibilidades de classificação, sem fornecer instâncias o suficiente para que seja possível distinguir os contextos entre conceitos.

Dentre os benefícios da utilização do KDC incluem-se a utilização de uma base de conhecimento, que separa o sistema de classificação do conhecimento empregado no processo, o que também permite aumentar a quantidade de informação utilizada para a classificação. Através da inserção de novos conceitos e instâncias, é possível expandir o vocabulário utilizado sem a necessidade de modificar o classificador ou retreinar algum modelo de classificação. Entretanto, esse processo deve ser acompanhado, pois a modificação na base de conhecimento pode impactar nos resultados obtidos pelo processo de classificação.

Além disso, a abordagem é independente do número de classes e de documentos a serem classificados, seu único requisito é que as classes devem estar presentes na base de conhecimento. Da forma como foi proposta, a abordagem também permite a restringir o conjunto de classes a serem usadas durante a classificação, e também definir diferentes formas para identificação de conceitos que representam classes.

Como limitações, a abordagem possui dependência de ferramentas de extração de termos e de uma base de conhecimento que possua uma função de relevância. A capacidade de extração de termos a partir de documentos, e o número de instâncias, conceitos e relacionamentos na base de conhecimento influenciam os resultados de classificação.

No decorrer do mestrado, foram geradas publicações visando divulgar, discutir e validar a proposta desenvolvida. Desta forma, foram publicados os artigos:

- DA SILVA, Gleidson Antônio Cardoso; DORNELES, Carina Friedrich. Descoberta de Domínio Conceitual de Páginas Web. Workshop on Thesis and Dissertations in Databases. SBBD 2014.

- DA SILVA, Gleidson Antônio Cardoso; DORNELES, Carina Friedrich. Towards Automatic Document Classification by Exploiting only Knowledge Resources. 34th International Conference of the Chilean Computer Science Society. SCCC 2015. Qualis B3.

Como trabalhos futuros, pretende-se estender o KDC para classificação com múltiplos rótulos, o que pode ser feito determinando-se uma regra para seleção de mais de um conceito ranqueado como classe para o documento. Outro ponto trata da realização de novas experiências com outras ferramentas

de extração de termos a partir de documentos e também outras bases de conhecimento. Tais experimentos podem ser usados para definir melhor as características desejáveis de ferramentas para uso em conjunto com o KDC por meio da análise de desempenho.

REFERÊNCIAS

AAS, K.; EIKVIL, L. **Text Categorization: A Survey**, Report No. 941. ISBN 82-539-0425-8, 1999.

AGGARWAL, C. C.; ZHAI, C. **A survey of text classification algorithms**, In Mining text data (pp. 163-222). Springer US, 2012.

ALBITAR, S.; ESPINASSE, B.; FOURNIER, S. **Towards a Supervised Rocchio-based Semantic Classification of Web Pages**, In KES (pp. 460-469), 2012.

ALBITAR, S.; FOURNIER, S.; ESPINASSE, B. **The impact of conceptualization on text classification**, In Web Information Systems Engineering-WISE (pp. 326-339). Springer Berlin Heidelberg, 2012.

ALLAHYARI, M.; KOCHUT, K. J.; JANIK, M. **Ontology-Based Text Classification into Dynamically Defined Topics**, In Semantic Computing (ICSC), IEEE International Conference on (pp. 273-278). IEEE, 2014.

BARLA; MICHAL; BIELIKOVA, M. **From Ambiguous Words to Key-Concept Extraction**, Database and Expert Systems Applications (DEXA),24th International Workshop on. IEEE, 2013.

BOLLACKER, K.; EVANS, C.; PARITOSH, P.; STURGE, T.; TAYLOR, J. **Freebase: a collaboratively created graph database for structuring human knowledge**, In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). ACM, 2008.

BRODIE, M. L.; JOHN, M. E. **On knowledge base management systems: integrating artificial intelligence and database technologies**, Springer Science & Business Media, 2012.

CHAKRABARTI, S.; AGRAWAL, B. D. R.; RAGHAVAN, P. **Using taxonomy, discriminants and signatures for navigating in text databases**, VLDB Conference, 1997.

CLARK, J.; KOPRINSKA, I.; POON, J. **A neural network based approach to automated e-mail classification**, In Web Intelligence, IEEE/WIC/ACM International Conference on (pp. 702-702). IEEE Computer Society. 2003.

DASGUPTA, A.; DRINEAS, P.; HARB, B. **Feature Selection Methods for Text Classification**, KDD'07, ACM, 2007.

DUMAIS, S.; CHEN, H. **Hierarchical classification of Web content**, SIGIR, 2000.

GAGNON; MICHEL; ZOUAQ, A.; JEAN-LOUIS, L. **Can we use linked data semantic annotators for the extraction of domain-relevant expressions?** Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

HARISH, B. S.; GURU, D. S.; MANJUNATH, S. **Representation and Classification of Text Documents: A Brief Review**, IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition RTIPPR, 2010.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA data mining software: an update**, ACM SIGKDD Explor Newslett, 11: 10–8, 2009.

HALEVY, A. Y.; ASHISH, N.; BITTON, D.; CAREY, M.; DRAPER, D.; POLLOCK, J.; SIKKA, V. **Enterprise information integration: successes, challenges and controversies**, In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, 778-787, 2005.

HE, WU; DA XU, LI. **Integration of distributed enterprise applications: a survey**, Industrial Informatics, IEEE Transactions on, v. 10, n. 1, 35-42, 2014.

HERNÁNDEZ, I., RIVERO, C. R., RUIZ, D., & CORCHUELO, R. CALA: **An unsupervised URL-based web page classification system**, Knowledge-Based Systems, 57, 168-180, 2014.

HOFFART, J. et al. YAGO2: **A spatially and temporally enhanced knowledge base from Wikipedia**, Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013.

HUANG, L.; MILNE, D.; FRANK, E.; WITTEN, I. H. **Learning a concept-based document similarity measure**, Journal of the American Society for Information Science and Technology, 63(8), 1593-1608, 2012.

HUSBY; STEPHANIE, D.; BARBOSA, D. **Topic classification of blog posts using distant supervision**, In Proceedings of the Workshop on Semantic Analysis in Social Media. Association for Computational Linguistics, 2012.

KO, Y., SEO, J. **Automatic text categorization by unsupervised learning**, 18th COLING, 1, 453-459, Association for Computational Linguistics. 20. 2000.

KORDE, V.; MAHENDER, C. N. **Text classification and classifiers: A survey**, International Journal of Artificial Intelligence & Applications (IJAIA), 3(2), 85-99. 2012.

LANG. K. **NEWSWEEDER: Learning to Filter Netnews**, ICML Conference, 1995.

LEHMANN, J. et al. **DBpedia-a large-scale, multilingual knowledge base extracted from Wikipedia**, Semantic Web Journal, v. 5, p. 1-29, 2014.

LEWIS, D. D. **An evaluation of phrasal and clustered representations on a text categorization task**, In Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 37-50). ACM. 1992.

LUCIA, W., FERRARI, E. EgoCentric: **Ego Networks for knowledge-based Short Text Classification**, CIKM, 1079-1088, ACM. 2014.

LUO, Q.; CHEN, E.; XIONG, H. **A semantic term weighting scheme for text categorization**, Expert Systems with Applications, 38(10), 1270812716. 2011.

MATERNA, J**. Automatic web page classification**, RASLAN, 8, 84-93. 2008.

PANG, G.; JIANG, S. **A generalized cluster centroid based classifier for text categorization**, Information Processing & Management, 49(2), 576-586. 2013.

PAK, A.; PAROUBEK, P. **Twitter as a Corpus for Sentiment Analysis and Opinion Mining**, In LREC (Vol. 10, pp. 1320-1326). 2010.

PORIA, S.; AGARWAL, B.; GELBUKH, A.; HUSSAIN, A.; HOWARD, N. **Dependency-based semantic parsing for concept-level text analysis**, In Computational Linguistics and Intelligent Text Processing (pp. 113-127). Springer Berlin Heidelberg, 2014.

QI, X.; DAVISON, B. D. **Web page classification: Features and algorithms**, ACM Computing Surveys (CSUR), 41(2), 12, 2009.

RECHENMANN, F. **Knowledge bases and computational molecular biology**, Towards Very Large Knowledge Bases, p. 7-12, 1995.

SAHAMI, M.; DUMAIS, S.; HECKERMAN, D.; HORVITZ, E. **A Bayesian approach to filtering junk e-mail**, In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105), 1998.

SALTON; GERARD; MCGILL, M. J. **Introduction to modern information retrieval**, 1983.

SHAPIRO, S. S.; WILK, M. B. **An analysis of variance test for normality (complete samples)**, Biometrika, 591-611, 1965.

SHEN, D.; CHEN, Z.; YANG, Q.; ZENG, H.; ZHANG, B.; LU, Y.; MA, W. **Web-page Classification through Summarization**, SIGIR, 2004.

SIGOGNE, A.; CONSTANT, M**. Real-time unsupervised classification of web documents**, IMCSIT'09, 281-286, IEEE, 2009.

SOUZA, A. M. **Análise de Variância ANOVA**, Departamento de Estatística. Disponível em <http://w3.ufsm.br/adriano/aulas/anova/T[0]anova.pdf>. Acesso em 16/11/2015.

STUDENT. **The probable error of a mean**, Biometrika, p. 1-25, 1908.

TAO, X.; LI, Y.; LAU, R. Y.; WANG, H. **Unsupervised multi-label text classification using a world knowledge ontology**, In Advances in Knowledge Discovery and Data Mining (pp. 480-492). Springer Berlin Heidelberg, 2012.

VIDAL, M. L.; DA SILVA, A. S.; DE MOURA, E. S.; CAVALCANTI, J. M. **Structure-Based Crawling in the Hidden Web**, J. UCS, 14(11), 1857-1876, 2008.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: practical machine learning tools and techniques**, 3.ed, Morgan Kaufmann, 2011.

XIAOGUANG, Q.; DAVISON, B. D. **Web page classification: Features and algorithms**, ACM Computing Surveys (CSUR), 2009.

YATES, R. B.; NETO, B. R. **Modern Information Retrieval**, 2a Edition, 2011.

YUN, J. et al. **Semantics-based representation model for multi-layer text classification**, In Knowledge-Based and Intelligent Information and Engineering Systems (pp. 1-10). Springer Berlin Heidelberg, 2010.

ZHAO, WEI; WANG, YAFEI; LI, DAN. **New Feature Selection Algorithm in Text Categorization**, International Symposium on Computer, Communication, Control and Automation, 2010.