# Classificação de textos baseada em ontologias de domínio*

Sandro J. Rigo[1,2], José Palazzo M. de Oliveira[1]
Cristiano Barbieri[2]

[1] Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
{sjrigo, palazzo}@inf.ufrgs.br
[2]Universidade do Vale do Rio dos Sinos – UNISINOS
barcristiano@gmail.com

**Abstract.** *With the enormous quantity of documents that are now available on the Web, accessing and collecting desired data has become a difficult task that produce low quality results. The text classification permits a reduction of this problem. This work aims to describe an implementation of a text classification system with the use of linguistic information, described in a domain ontology with the necessary information for the identification of both structure and concepts of documents associated to a specific class.*

**Resumo.** *Com a enorme quantidade de documentos disponíveis na Web, o acesso aos dados desejados tornou-se uma tarefa difícil e que gera resultados de baixa qualidade. A classificação de textos permite uma redução deste problema. Este trabalho propõe-se a realizar classificação de textos com uso de informações lingüísticas, descritas em uma ontologia de domínio contendo as informações necessárias para a identificação da estrutura e conceitos dos documentos associados a uma classe específica.*

## 1. Introdução

A Web apresenta-se como um grande repositório de informações heterogêneas. Ao mesmo tempo, observa-se o aumento na quantidade de documentos disponíveis para consulta. Como conseqüência, uma das grandes características da Web é a sobrecarga de informações. Segundo Cumbo et al (2004), administrar a quantidade enorme de documentos textuais disponíveis na Internet tornou-se um problema importante de administração do conhecimento. Para isso, faz-se necessário um sistema que possua mecanismos efetivos para classificar a informação contida nestes documentos textuais. O projeto Rec-Semântica tem por objetivo indicar de forma automática a um usuário o conteúdo que pode ser relevante ou interessante a ele quando do acesso a um servidor Web ou da utilização de um sistema cujo principal objetivo é o trabalho sobre Informação. Nesta categoria encontram-se sistemas de recuperação de informação, de apoio ao processo de

* Este artigo está inserido no projeto REC-SEMANTICA, Plataforma de Recomendação e Consulta na Web Semântica Evolutiva, PRONEX – FAPERGS/CNPq.

decisão, de ensino suportado por computadores, entre outros. O projeto aborda o problema de recomendação sob vários aspectos correlacionados. Em primeiro lugar, é proposta a utilização de uma abordagem semântica, isto é de uma abordagem baseada em ontologias de domínio, tanto para modelar os perfis de usuários, quanto para classificar conteúdos a serem avaliados para recomendação. Dentro deste enfoque a categorização de textos, com base na semântica associada é de relevante importância.

A classificação ou categorização de textos é uma das técnicas que possibilitam auxílio na localização de resultados desejados em pesquisas e em sistemas de recomendação, minimizando a sobrecarga de informação. Resumidamente, consiste no processo de classificar automaticamente um conjunto de documentos em uma ou mais categorias pré-existentes facilitando a busca seletiva de informações. Segundo Peixoto et al (2006), o processo de categorização de textos, ou classificação automática de documentos, proporciona uma resposta para a necessidade de se separar a informação em categorias, que facilitem a respectiva manipulação e recuperação. Uma das abordagens gerais observadas em grande número de trabalhos de classificação de textos consiste no uso de técnicas de aprendizado de máquina baseadas em conjuntos de dados de treinamento. Nestes trabalhos são utilizados exemplos de documentos e a indicação de resultados esperados para cada conjunto específico. O algoritmo de aprendizado deve ser capaz de reproduzir os resultados a partir destes exemplos. Algumas abordagens utilizadas são árvores de decisão, redes neurais artificiais, classificadores *bayesianos* e *suport vector machines* (Silva et al, 2004b).

Entretanto, estas abordagens podem apresentar deficiências no tratamento de problemas comuns à área de recuperação de informações, como o tratamento de sinonímia ou polissemia. Em função disso, alguns trabalhos utilizam recursos lingüísticos e semânticos, como forma de minimizar estes problemas. Alguns tratam de informações lingüísticas na etapa de pré-processamento de textos para a classificação, realizada por intermédio de redes neurais (Silva et al, 2004a). Outros tratam a classificação de textos a partir de ontologias de domínio que são automaticamente adquiridas por intermédio de regras morfológicas e métodos estatísticos, como em Wu et al (2003). Outro exemplo pode ser observado na classificação de textos utilizando ontologias de domínio na área médica, sendo as mesmas utilizadas para aprendizado automático sobre os textos analisados (Bloehdorn et al, 2005). Ainda podem ser encontrados trabalhos de classificação de textos baseados na hierarquia semântica das palavras e relações entre grupos de conceitos e significados, como no exemplo de Peng et al (2005), onde é utilizada grande quantidade de processamento sobre reconhecimento de palavras que possuam relacionamento. Ou, como no trabalho apresentado por Cumbo et al (2004), trata-se a classificação de textos com a utilização de ontologias de domínio e uso de linguagem de programação em lógica *Datalog*, onde, através da criação de predicados determina-se em qual domínio específico um determinado documento pode ser classificado.

Mesmo nos trabalhos com uso de ontologias ou informações lingüísticas observa-se o mesmo procedimento geral adotado nos trabalhos citados anteriormente, ou seja, a partir de um conjunto de documentos de treinamento são adicionadas informações semânticas e o resultado é tratado por um algoritmo de aprendizado de máquina, sendo o resultado utilizado para posterior classificação de novos documentos. O presente trabalho propõe-se a realizar classificação de textos com uso de informações lingüísticas, descritas em uma ontologia de domínio contendo as informações necessárias para iden-

tificação da estrutura e conceitos dos documentos, associados a uma classe específica. O processo desenvolvido é similar ao processo de classificação de textos, envolvendo inicialmente o tratamento de um conjunto de documentos em domínio de conhecimento específico para as avaliações necessárias. Com as informações obtidas desta análise, são realizadas operações para a obtenção de informações estatísticas e utilizadas heurísticas para extrair construções que possam ser relevantes na identificação de documentos do domínio, a partir da descrição de frases e de informações morfossintáticas. Este resultado será utilizado em consultas à ontologia de domínio, para oportunizar diferentes comparações em relação ao documento analisado.

A seguir é descrito o trabalho proposto, sendo que na seção dois são apresentados os conceitos de classificação de textos, além de uma análise sobre trabalhos relacionados. Na seção três é descrita a metodologia geral adotada e a ontologia criada. Na seção quatro são descritos experimentos realizados e por fim, na seção cinco são descritas as conclusões do trabalho.

## 2. Classificação de textos e trabalhos relacionados

A classificação de textos permite a identificação automática dos mesmos, verificando se o conteúdo de cada documento em um determinado conjunto pertence ou não a um domínio estabelecido. Existem diversos tipos de classificadores de textos automáticos, sendo que em geral as diferentes categorias são criadas a partir dos termos mais freqüentes encontrados nos textos. Isso pode gerar problemas relacionados a termos polissêmicos ou em casos de sinonimia. Outra metodologia para a classificação de textos baseia-se na utilização de informações lingüísticas. Nesta abordagem podem ser utilizadas diversas das técnicas conhecidas, com o acréscimo de informações complementares. Para tanto é necessária a realização de transformações sobre os documentos a serem classificados, a fim de que se obtenha uma representação estruturada envolvendo, conforme Silva et al (2004a), a remoção de termos irrelevantes, a construção da estrutura sintática, evitando o volume demasiado grande de informações gramaticais e a extração de construções relevantes, para a utilização no processo de classificação de textos.

Conforme Bloehdorn et al (2005), podem ser observadas melhorias em tarefas de classificação de textos com uso das características conceituais extraídas de ontologias. A abordagem é baseada na distribuição de hipóteses, ou seja, verifica-se, durante o processo de classificação, se os termos são semanticamente similares ao contexto ao qual eles estão compartilhados. Utilizam-se das seguintes características: conceito de anotação e generalização. A anotação possui a finalidade de evitar consultas desnecessárias à ontologia. A generalização tem por finalidade adicionar conceitos mais gerais aos conceitos mais específicos encontrados nos textos. Desta forma, existe uma abrangência maior relacionada com os conceitos e aumenta-se a similaridade entre diversos documentos analisados, que possuam características em comum. Para tanto utiliza-se ontologias de domínio que foram extraídas do corpus de textos.

O artigo de Wu et al (2003) trata da categorização de textos baseada em ontologias de domínio. Estas são adquiridas através de regras morfológicas e métodos estatísticos. É uma forma de criar e melhorar ontologias conforme se adquire conhecimento. Uma ontologia de domínio pode ser utilizada para identificar a estrutura conceitual das sentenças em um documento e, além disso, esta ontologia pode ser utilizada como base

para diversas aplicações, como por exemplo: aplicações de perguntas e respostas, de gerenciamento do conhecimento e de organização da memória. Segundo os autores, a vantagem em se utilizar ontologias, comparada com outros mecanismos de representação do conhecimento é que a mesma pode ser lida, intepretada e editada por seres humanos. Erros podem ser detectados e com isso a descrição pode ser melhorada. Outra vantagem é a possibilidade do compartilhamento da ontologia por várias aplicações.

No artigo de Silva et al (2004a) está descrita uma avaliação da utilização de informação lingüística para categorização da informação de textos da Língua Portuguesa, através do uso de redes neurais. O processo de categorização é composto por três etapas: coleta de base, onde é feita a busca de documentos relevantes ao domínio da aplicação para ser extraído o conhecimento; pré – processamento, com remoção de termos irrelevantes (*stopwords*), redução de afixos e seleção de termos; categorização: realizada através do aprendizado de máquina, com uso de rede neural artificial. O método de classificação proposto no trabalho necessita uma grande quantidade de documentos para o treinamento da rede neural usada na classificação de textos. Através de um analisador sintático denominado PALAVRAS Xtractor[1] os documentos, incialmente em formato de linguagem natural, são convertidos em três documentos XML[2] (*eXtensible Markup Language*): *Words*, *POS* e *Chunks* (descritos adiante no texto). A extração de combinações gramaticais é realizada utilizando folhas de estilo sobre os arquivos XML gerados. A partir desta extração, são geradas listas de termos resultantes para a fase de categorização, sendo empregada para isto a técnica de aprendizado de máquina. Os termos mais relevantes foram identificados no conjunto de treino com base no cálculo de freqüência relativa. Esta relação de termos resultantes constituiu o vetor local de cada categoria. Baseado nos vetores locais das categorias foi construído um vetor global adotando a representação de documentos do modelo de espaço vetorial baseado na categorização múltipla dos exemplos. Uma vez concluída a etapa de construção dos vetores globais, estes foram submetidos à ferramenta Weka[3] (*Waikato Environment for Knowledge Analysis*), para o treinamento da rede neural artificial MLP (*Multi-layer Perceptron*), utilizando o algoritmo BP (*Backpropagation*) para o aprendizado.

Alguns trabalhos fazem uso de classificação de textos utilizando ontologias e regras para categorização dos mesmos. Em Cumbo et al (2004), é utilizada a combinação de linguagens de programação em lógica com ontologias de domínio, para a classificação de textos. A linguagem utilizada, *Datalog*, proporciona a manipulação da semântica fornecida pela ontologia de domínio. O mesmo descreve padrões complexos, utilizados na classificação de textos da seguinte forma: para cada texto é feita uma verificação das regras e de acordo com um percentual de regras válidas estipulado, considera-se o texto como pertencente a uma categoria. A idéia básica de utilização da programação em lógica é realizar o reconhecimento dos conceitos dentro dos textos. Realiza-se a categorização baseando-se em dois tipos de predicados pré-definidos: predicados de pré-processamento, onde se estipulam as regras de manipulação sobre os textos dos do-

[1] *http://visl.sdu.dk/visl/pt/parsing/automatic/*
[2] *http://www.w3.org/XML/*
[3] *http://www.cs.waikato.ac.nz/ml/weka*
[4]

cumentos analisados e os predicados de ontologia, que representam as regras sobre as ontologias de domínio, como por exemplo: associações ou sinonímia.

Desta forma, pode ser destacado que existe uma tendência ao uso de recursos semânticos e lingüísticos em tarefas de classificação de textos. Os trabalhos comentados possuem diversos elementos interessantes para o classificador desenvolvido, sendo que apontam para a possibilidade de melhoria de resultados com o uso destes recursos adicionais, ultrapassando deste modo a limitação do tratamento dos textos com uso apenas da sintaxe.

## 3. Metodologia desenvolvida

Nesta seção são descritas e discutidas as etapas para a construção do classificador de textos. De forma bastante resumida, o processo inicia com avaliações de *corpus* em diferentes documentos, neste caso relacionados ao domínio de linguagens de programação OO (Orientada a Objetos). A partir destas avaliações, feitas por especialistas no domínio em questão, obtém-se elementos para o desenvolvimento da ontologia de domínio a ser empregada. Em seguida são realizadas as análises léxica e sintática sobre diversos documentos a serem classificados, sendo que os resultados destas análises são arquivos gerados em formato XML. Estes arquivos contém informações morfo-sintáticas e livres de termos irrelevantes, para serem utilizados diretamente pelo classificador de textos. Com estes documentos manipulados e a ontologia desenvolvida, serão realizadas as consultas, com o objetivo de classificação de diversos textos. Estas consultas permitem, junto com heurísticas, a delimitação dos elementos de interesse, dentro dos documentos, para a escolha daqueles pertencentes a esta classe específica.

A avaliação de corpus possui por finalidade obter como resultado a classificação de textos restritos a Língua Portuguesa e em especial relacionados às linguagens de programação OO. Para tanto, inicialmente foi realizado estudo sobre um conjunto de textos em formato HTML5 (Hyper Text Markup Language), nos quais identificaram-se os principais conceitos encontrados em um manual de linguagem de programação, tais como: objeto, herança, classes, métodos, atributos, entre outras mais, que estão diretamente relacionadas com o estudo do corpus em questão. Segue abaixo um pequeno trecho de texto de um dos documentos analisados (CESTA et al ,1996a), onde observa-se em negrito o conjunto de conceitos de domínio identificados.

> *"Uma **classe** é um tipo definido pelo usuário que contém o molde, a especificação para os **objetos**, algo mais ou menos como o tipo inteiro contém o molde para as variáveis declaradas como inteiros. A **classe** envolve, associa, funções e dados, controlando o acesso a estes. Definí-la implica em especificar os seus **atributos** (dados) e seus **métodos** (funções).*"

Através de estudos realizados sobre um conjunto de documentos, obteve-se como resultado a descrição dos termos mais freqüentes. Além desta característica são identificados sinônimos e as possíveis hierarquias entre os conceitos. Todas estas informações detectadas durante os estudos de *corpus* foram mapeadas manualmente em uma ontologia de domínio com a descrição desta estruturação, bem como as relações impor-

[5] *http://www.w3.org/html/*
[7] *http://protege.stanford.edu*

tantes identificadas. A ontologia de domínio foi criada com conceitos visando mapear dois tipos distintos: conceitos sobre a estrutura do documento e sobre o próprio domínio em questão. Os conceitos de estrutura permitem identificar se o documento analisado possui realmente as características que organizam um documento de linguagem de programação OO. No caso do experimento realizado, os documentos analisados são tutoriais compostos, por exemplo, com as seguintes partes: índice, introdução, bibliografia, entre outros. De forma semelhante, os conceitos de domínio são empregados na identificação de características do conteúdo relacionadas com o foco utilizado na experiência inicial, ou seja, linguagens de programação OO. Alguns exemplos destes conceitos já foram mencionados anteriormente no texto. Para a criação desta ontologia de domínio, utilizou-se a ferramenta *Protégé*[7]. Na Figura 1, tem-se uma visão parcial dos conceitos descritos na ontologia de domínio criada a partir dos textos analisados.


Figura 1: Visão dos conceitos no ambiente do *Protégé*.

Dentre os conceitos visualizados na figura 1, onze são pertencentes ao domínio estrutural de um tutorial de linguagem de programação OO. São eles: conceito, configuração, fundamentos, guia, índice, instalação, introdução, manual, prefácio, sumário e tutorial. Os outros quatro conceitos referentes ao domínio conceitual são: atributo, classe, método e objeto. Após a definição dos conceitos da ontologia de domínio foi realizada a criação de sinônimos e relacionamentos de hierarquia entre os mesmos. Para tanto, criaram-se propriedades definindo-se para as mesmas os tipos sobre os quais cada uma irá exercer determinada função entre os conceitos da ontologia de domínio. Estas propriedades permitem consultas que identificam os termos como pertencentes ou não à ontologia. Caso pertençam, permitem identificar seu tipo (conceitos do domínio ou partes da estrutura). No caso de sinônimos, estes mesmos elementos da ontologia permitem a sua identificação, a partir de consultas mais específicas.

A figura 2 ilustra o processo geral proposto para a classificação de textos. O processo de classificação dos documentos inicia com o pré-processamento do conteúdo dos mesmos, através do *parser* PALAVRAS, um analisador sintático para a Língua Portuguesa, desenvolvido junto ao projeto VISL (*Visual Interactive Syntax Learning*). Após a etapa de pré-processamento é realizada a análise sintática dos termos resultantes. Para a realização da mesma, utiliza-se a ferramenta PALAVRAS Xtractor, que recebe o resultado do *parser* PALAVRAS e gera três novos arquivos XML. O primeiro arquivo XML gerado, de nome *words*, contém todos os termos relevantes identificados por um atributo de nome *id*, possuindo valores únicos. O segundo arquivo XML gerado, de nome

POS, contém as informações morfo-sintáticas e a forma radical de cada um das palavras existentes no arquivo de *words*. O terceiro e último arquivo XML gerado é o de *chunks*, contendo as estruturas sintáticas das sentenças definidas no arquivo de POS. O arquivo de *chunks* possui entre outras características o elemento *paragraph*, identificando cada um dos parágrafos do texto analisado e atribuindo um identificador único ao mesmo, de nome *id.* Além disso, para cada frase deste parágrafo um elemento de nome *sentence* é definido e, assim como *paragraph*, também é identificada por um valor único.

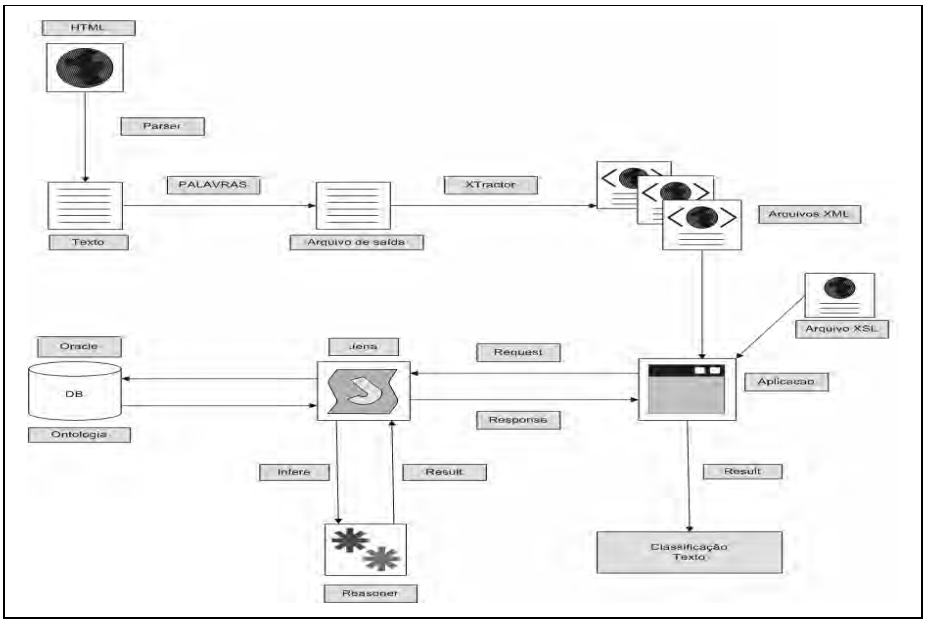


Figura 2: Arquitetura do classificador de textos proposto.

A partir destes arquivos XML são extraídos os termos simples ou compostos a partir das combinações gramaticais consideradas relevantes para o processo de classificação de textos. Estas extrações são realizadas com folhas de estilo, nas quais são especificadas as principais regras para obtenção dos resultados desejados. Após esta extração, é gerada uma lista de termos resultantes, sendo a mesma utilizada em consultas realizadas sob a ontologia de domínio. A linguagem SPARQL[8] (*Protocol for access RDF and Query Language*) foi utilizada para recuperar as informações na ontologia de domínio. A biblioteca *Jena*[9] possibilita duas formas para obtenção de tais informações: uma é através do mecanismo de inferência, utilizando-se o seu próprio mecanismo de inferência ou mediante ao uso de um componente externo via DIG (*Description Logic Reasoner Interface*). Neste trabalho optou-se pela primeira forma.

Além disso, foi realizada a integração da biblioteca *Jena* com um Banco de Dados Oracle[10] para que a ontologia de domínio fosse armazenada no mesmo, em forma de triplas RDF[11] (*Resource Description Framework*). Para cada consulta executada a onto-

[8] http://www.w3.org/TR/rdf-sparql-query/
[9] http://jena.sourceforge.net/
[10] http://www.oracle.com
[11] http://www.w3.org/RDF/

logia de domínio, utilizam-se os resultados fornecidos pelas mesmas na realização de operações, obtendo-se, com estas, informações estatísticas. Ao mesmo tempo, com a utilização de heurísticas e da ontologia identificam-se construções relevantes para a classificação de documentos do domínio e comparações entre os documentos analisados. Antes mesmo de classificar um texto é necessário que se crie um arquivo XSL (*eXtensible Stylesheet Language*)[12], contendo as regras necessárias para obter através do mesmo as construções relevantes a serem utilizadas nas consultas a ontologia de domínio. Após obter o resultado da consulta são realizadas contabilizações para a obtenção das seguintes informações: número total de termos processados (independentemente de repetições), número total de termos mapeados em conceitos na ontologia de domínio, quantidade de documentos processados, quantidade de termos encontrados pro elemento pesquisado.

## 4. Experimentos realizados

Foram realizados experimentos com textos de diferentes tipos de domínios. Todos os documentos disponíveis para o classificador de textos são textos anotados com informações sintáticas e no formato XML, ou seja: *words*, POS e *chunks*. Para os experimentos iniciais utilizaram-se apenas os arquivos de *words*, contendo todas as palavras dos textos. O conjunto de documentos é constituído de 46 textos, sendo todos pertencentes a Língua Portuguesa. Os textos se encontram distribuídos conforme apresentado a seguir. Dois textos[13,14] referentes a tutoriais de linguagem de programação OO, de onde foram extraídas partes de seu conteúdo e feita a anotação dos mesmos. Quatro textos jornalísticos sobre o cotidiano da Folha de São Paulo[15]. Sete textos do livro didático de Ciências da Editora Scipione, sobre Corpo Humano, do ano de 1992. Os temas dos arquivos analisados são respectivamente: Crescimento, Células e Hereditariedade, Organização do Nosso Corpo e Funções vitais[16]. Trinta e três textos jornalísticos da Folha de São Paulo, ano de 1994, seções: Esporte, Imóveis, Informática, Política e Turismo[6]. A tabela 1 ilustra os resultados obtidos.

Tabela 1. Resultados dos experimentos

| Experimento | Tipo do documento |
|---|---|
| 1 | Textos jornalísticos sobre o cotidiano. |
| 2 | Textos jornalísticos sobre esportes. |
| 3 | Textos jornalísticos sobre imóveis. |
| 4 | Textos jornalísticos sobre informática. |
| 5 | Textos jornalísticos sobre política. |
| 6 | Textos jornalísticos sobre turismo. |
| 7 | Textos didáticos sobre ciência. |
| 8 | Tutoriais de linguagens de programação OO. |

| Avaliação\Experimentos | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Qtd. de docs. processados | 4 | 6 | 6 | 6 | 6 | 9 | 7 | 2 |
| Qtd. de palavras processadas | 1.288 | 1.356 | 1.672 | 959 | 1.876 | 2.596 | 15.470 | 8.415 |
| Qtd. de ocorrências de term. mapeados | 0 | 0 | 2 | 0 | 4 | 6 | 1 | 218 |
| Qtd. de termos mapeados em conceitos de domínio | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 199 |
| Qtd. de termos mapeados em conceitos de estrutura | 0 | 0 | 0 | 0 | 4 | 3 | 1 | 19 |

[12] http://www.w3.org/Style/XSL/
[13] *http://www.ic.unicamp.br/~cmrubira/aacesta/java/javatut.html*
[14] *http://www.ic.unicamp.br/~cmrubira/aacesta/cpp/cpp15.html*
[15] *http://www.inf.unisinos.br/~renata/laboratorio/mais_jd_mc.htm*
[16] *http://www.inf.unisinos.br/~renata/laboratorio/mais_jd_mc.htm*

Todos os experimentos foram realizados com uma quantidade pequena de documentos, além dos textos possuírem também uma quantidade bastante pequena de palavras, se comparadas com outras experiências de classificação de textos. Mesmo com as características mencionadas, os resultados obtidos nos experimentos demonstram que a ontologia de domínio gerada, com apenas 15 conceitos, sendo 11 destes pertencentes ao domínio de estrutura e os restantes pertencentes ao domínio conceitual, como já mencionado, permite ao método de classificação implementado distinguir os textos pertencentes a tutoriais de linguagens de programação OO, dos demais textos avaliados.

Tabela 2. Percentuais dos tipos de conceitos mapeados em cada experimento

| Experimento | Conceitos de Domínio % | Conceitos de Estrutura % |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 0 |
| 3 | 25 | 0 |
| 4 | 0 | 0 |
| 5 | 0 | 9,09 |
| 6 | 25 | 9,09 |
| 7 | 0 | 9,09 |
| 8 | 100 | 27,27 |

Avaliando a Tabela 2, que apresenta o comparativo dos valores percentuais mapeados em relação aos tipos de conceitos na ontologia de domínio, torna-se mais evidente que nos experimentos realizados somente o de número 8 obteve 100% de êxito no que se refere ao domínio conceitual de uma linguagem de programação OO.

## 5. Conclusões

Este trabalho apresentou uma proposta para classificação de textos utilizando informações linguísticas e ontologias de domínio, com a finalidade de melhorar o processo de classificação para um domínio específico. Esta proposta segue uma tendência de utilização de ontologias de domínio como forma de auxiliar na classificação de textos e utilização das mesma para refinar o conhecimento adquirido. Alguns trabalhos utilizam informações linguísticas no processo de classificação, mas com limitações quanto ao modelo proposto, referente à forma de aquisição do conhecimento. Já em outros trabalhos, além de utilização de ontologias de domínio como base do conhecimento, observa-se o uso de recursos de inferência para classificação dos textos.

Através do estudo de diversos textos de tutoriais de linguagens de programação OO, foi definido um pequeno conjunto de conceitos de estrutura e de domínio. Além disso, foi utilizada a informação de 46 textos anotados, compreendendo diferentes domínios. Estes textos se encontram na forma anotada em três arquivos XML, contendo sentenças válidas para serem consultadas na ontologia de domínio.

Pretende-se realizar a utilização de conceitos com nomes compostos e construir relações entre conceitos que descrevam frases e identifiquem o contexto em que se encontram as mesmas, como por exempo: *Uma classe implementa um tipo abstrato de dados*. A partir desta frase é possível implementar consultas à ontologia de domínio, verificando-se a existência de conceitos com uma estrutura semelhante. Pode-se também aumentar a especialização dos conceitos de domínio descritos na ontologia, acrescentando mais identificadores, referentes a linguagens de programação OO, como exemplo:

construtor, herança, instância, polimorfismo, entre outros. Além disso, para que possam ser realizadas consultas em SPARQL mais complexas, conforme os exemplos acima mencionados, é essencial a utilização dos arquivos de POS e *chunks*, contendo informações morfo-sintáticas mais detalhadas sobre os textos anotados.

Este trabalho constitui-se em uma parcela de um projeto de pesquisa de maior escopo que permitirá a recomendação de artigos científicos ou técnicos na área da informática levando em conta os conceitos expressos nos documentos e o perfil do usuário. Em outra atividade deste projeto (Lopes, 2007) demonstrou a factibilidade de realizar a recomendação de artigos com base no perfil do usuário identificado através de suas publicações. Esta validação permite afirmar que a classificação de textos com base em otologias é uma ferramenta muito poderosa para o suporte a mecanismos adaptáveis de recomendação.

## Referências

Bloehdorn, S.; Cimiano, P. ; Hotho, A.. Learning Ontologies to Improve Text Clustering and Classification. Proceedings of the 29th Annual Conference of the German Classification Society (GfKl 2005), Magdeburg, Germany, March 9-11, 2005, volume 30 of Studies in Classification, Data Analysis, and Knowledge Organization, pp. 334-341. Springer, February 2006.

CESTA, A. Augusto; RUBIRA, C. M. F. (1996a). A linguagem de programação Java. Disponível em: http://www.ic.unicamp.br/~cmrubira/aacesta/java/javatut.html acessado em (02/11/2006).

Cumbo, Chiara; Iritano, Salvatore; Rullo, Pasquale. Combining logic programming and domain ontologies for text classification. CILC 2004, Convegno Italiano di Logica Computazionale. 16-17 giugno 2004, Dipartimento di Matematica, Università di Parma. 2004.

Lopes G.R., Souto M.A.M., Wives L.K., Palazzo M. de Oliveira J.. Personalizing Bibliographic Recommendation under Semantic Web Perspective, in International Workshop on Web Information Systems Modeling (WISM 2007), held in conjunction with CAiSE 2007, June 12, 2007, Trondheim, Norway

Peixoto, F. D. Maria, Batista, Maria, Capeló, Maria. Categorização de textos. Departamento de Informática da Universidade da Beira Interior.

Peng, X., Ben Choi. Document Classifications Based on Word Semantic Hierarchies. Proceedings: Artificial Intelligence and Applications - 2005. Innsbruck, Austria.

Silva C., Vieira R., Osório, F. Uso de Informações Lingüísticas em Categorização de Textos utilizando Redes Neurais Artificiais. In: VIII Simpósio Brasileiro de Redes Neurais, (2004a), p. 1-6.

Silva, C.; Vieira, R.; Osório, F.S.(2004b), Uso de Informações Lingüísticas na etapa de pré-processamento em Mineração de Textos. II Workshop de Teses e Dissertações em Inteligência Artificial WTDIA'2004.

Wu, S.-H.; Tsai, Tzong-Han and Hsu, W.-L. Text Categorization Using Automatically Acquired Domain Ontology. the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL-03). Sapporo: Japan. 2003.