# Classificação Contínua de Documentos com Vocabulários Temáticos Dinâmicos para a Desambiguação de Termos

**Adriano A. Santos[1], Ulrich Schiel[2]**

[1]Programa de Pós-Graduação da Universidade Federal de Campina Grande (UFCG) – Paraíba – Brasil

[2]Departamento de Sistemas e Computação – Universidade Federal de Campina Grande (UFCG) – Paraíba - Brasil

adriano@copin.ufcg.edu.br, ulrich@dsc.ufcg.edu.br

***Abstract:*** *Get accurate results that reflect the desire of users is one of the main purposes of Information Retrieval Systems. A previous thematic classification of the documents to be retrieved is an interesting technique to get a good term disambiguation and consequently a better precision in the retrieval process. However, traditional techniques for classifying documents use static universal knowledge bases and cannot represent the dynamics of linguistic knowledge evolution and, even less, the spatial-temporal dependency of the meaning of words. Based on that, the purpose of this paper is to present a technique that considers the dynamics of word meanings through a context sensitive Thesaurus.*

***Resumo:*** *Obter resultados precisos que reflitam o desejo dos usuários é um dos maiores desafios dos Sistemas de Recuperação de Informações. Uma prévia classificação temática dos documentos permite uma desambiguação prévia dos termos e garante maior precisão na recuperação de documentos. No entanto, as técnicas tradicionais de classificação de documentos utilizam bases de conhecimento estáticas e universais, que não representam a dinâmica da evolução do conhecimento linguístico e, muito menos, a dependência espaço-temporal do sentido das palavras. Com base nisso, a proposta deste trabalho é apresentar uma técnica que considera esta dinâmica por meio de um Thesaurus sensível ao contexto.*

## 1. Introdução

O desenvolvimento da Internet possibilitou, entre outros aspectos, a disponibilização de uma vasta quantidade de documentos de fácil acesso [Nunes 2009], fazendo com que, devido ao número de domínios de conhecimento diferentes existentes, gerasse ambiguidade entre termos comuns entre esses domínios. Com isso, os usuários de Sistemas de Recuperação da

Informação (SRI) têm que realizar refinamento nos resultados de uma consulta realizada, para obter o desejado.

Técnicas convencionais de recuperação de informações não são adequadas para o processo de desambiguação do sentido das palavras por basearem-se em valores léxicos, sintáticos e análise estatística de texto [Tristão *et al* 2004]. Essas abordagens não vão além da aparência superficial das palavras.

Para amenizar a responsabilidade atribuída ao usuário de classificar os resultados obtidos por meio de buscas, foram criados sistemas para organização do conhecimento [Tristão *et al* 2004] que incluem uma variedade de esquemas que organizam, gerenciam e recuperam informações.

Esses sistemas abrangem o uso de 1) esquemas de classificação, [Navigli 2009] que são conjuntos de conceitos organizados sistematicamente de acordo com os critérios ou características escolhidas, 2) tesauros, [Tristão *et al* 2004] definidos como um vocabulário de termos relacionados semanticamente sobre determinada área de conhecimento e 3) ontologias, [Navigli 2009] usados como modelo de identificar os conceitos relevantes, dado um determinado fenômeno.

No processo de classificação de documentos podemos necessitar da busca do sentido das palavras existentes no documento para que seja possível uma classificação mais eficaz dos mesmos. Isso porque, conforme já mencionado, algumas palavras podem ser ambíguas e prover sentidos diferentes, o que dificulta uma classificação correta.

Segundo Adam Kilgarriff [Agirre e Edmonds 2007], o problema de desambiguar o sentido de uma palavra é a variedade de sentidos que uma palavra pode ter, já que cada palavra está associada a um contexto (relação entre o texto e a situação em que ele ocorre) no qual foi empregada e nem sempre identificar esse contexto é uma atividade trivial. Os sentidos de uma palavra, geralmente, estão registrados em um dicionário [Agirre e Edmonds 2007]. Essa fonte deve ser precisa, confiável e abrangente.

Esse problema é descrito como um problema IA-Completo [Mallery 1988] (uma analogia aos problemas NP-Completo) e é considerado de complexidade equivalente aos problemas mais difíceis em Inteligência Artificial [Navigli 2009], pois busca extrair semântica dos textos com processamento de linguagem natural e, na maioria das vezes, de forma automática e sem intervenção humana.

Muitas vezes, os SRI não conseguem encontrar documentos importantes devido às diferenças entre o vocabulário encontrado no processo de indexação de documentos com o da consulta e, também, porque o sentido das palavras pode mudar com o tempo.

Isso significa dizer que o sentido de uma palavra é contextual e, de acordo com as mudanças provocadas pelo tempo e espaço, determinadas classificações podem ser inadequadas. Além disso, novos termos surgem continuamente que ainda não foram devidamente classificados.

O processo de criação de novos termos pode ser exemplificado com o termo "Computação nas Nuvens". Se há alguns anos atrás, um determinado usuário realizasse uma pesquisa por este termo em um engenho de busca, os resultados obtidos tratariam de

programas de meteorologia. O termo "Computação nas Nuvens" no domínio do conhecimento da "Computação" é relativamente novo e vem se tornando mais expressivo com o surgimento de mais documentos sobre este tema na Internet. Por outro lado, o termo "rapariga" certamente terá um retorno diferente em um SRI em Portugal e no Brasil.

A partir dos exemplos expostos é possível perceber que o valor ou a importância de um conceito é variável de acordo com o tempo e local. Também, é possível perceber que o sentido das palavras é algo dinâmico e que está restritamente relacionado ao contexto e ao tempo. Uma palavra em inglês pode ter um significado nos Estados Unidos e outro na Inglaterra. O significado de um termo também pode ser diferente no meio acadêmico do que no meio jornalístico.

Considerando situações desse tipo, podemos deduzir que os sistemas de classificação de documentos tradicionais, por utilizarem bases de conhecimento universais e estáticas, não levam em consideração o contexto do sentido das palavras. Sendo assim, devemos encontrar uma técnica que possibilite considerar a dinâmica e dependência contextual do processo de aquisição de novos termos e atualização de conceitos consagrados, promovendo melhor acurácia nos resultados de um SRI.

Com base no exposto é proposto, neste trabalho, o desenvolvimento de um classificador de documentos que utilizará vocabulários temáticos obtidos a partir de documentos da Wikipédia para a classificação de documentos considerando contextos temporais e espaciais. O idioma escolhido para o experimento será o inglês por se tratar do idioma que tem servido como base em várias pesquisas desta área, possibilitando a comparação dos resultados.

Um vocabulário temático é formado por um conjunto de termos específicos (ou típicos) de um domínio de conhecimento. Cada domínio existente no experimento será representado por um vetor de termos, contendo o vocabulário temático do domínio com informações de frequência. Estes vetores serão comparados com os vetores extraídos dos novos documentos para classificá-los (técnica supervisionada). Os termos serão extraídos com o uso de *Part-of-speech Tagging* dos documentos e a comparação dos vetores utilizará a distância Euclidiana entre eles.

Deseja-se que o processo de classificação de documentos seja dinâmico no sentido de que, a cada novo termo significativo encontrado no documento que não esteja presente no vocabulário temático ao qual o documento foi classificado, o sistema realizará a atualização deste domínio com o novo termo. Porém, para que isso ocorra, será necessária uma busca em fontes externas ao sistema para garantir a importância do termo para o domínio. Com base nesse processo pretende-se obter uma melhor classificação dos documentos, tornando o processo de atualização dos termos continuo.

Para representar o contexto do sentido das palavras e as associações linguísticas dos termos é proposto o uso de um *Thesaurus* que leve em consideração as dependências contextuais dos termos registrados. A temporalidade do sentido das palavras, também estará representada no *Thesaurus*. Cada novo termo identificado como pertencente a um domínio será inserido no vocabulário temático e serão registradas as informações espaço-temporais.

Com esta informação será possível determinar o período da criação e classificação deste novo termo em um domínio de conhecimento.

O objetivo deste trabalho é desenvolver um classificador de documentos por domínios considerando contextos como base para o processo de desambiguação do sentido das palavras.

Durante o processo de indexação do documento serão identificados novos termos com seus devidos enquadramentos contextuais. Serão utilizadas técnicas de aquisição automática de conhecimento por meio de associações linguísticas com o uso de *Thesaurus*, levando em consideração o significado espaço-temporal do sentido das palavras.

## 2. Revisão Teórica e Trabalhos Relacionados

Em linhas gerais, podemos definir as abordagens utilizadas no processo de Desambiguação do Sentido da Palavra em [Agirre e Edmonds 2007] a) Base de Conhecimento, utilizam dicionários para classificação das palavras; b) Não-supervisionados, [Gliozzo *et al* 2004] evitam completamente as informações externas e trabalham diretamente com as informações da corpora; c) Supervisionado; utilização de marcações realizadas por especialista e d) Combinadas; utilizam combinações entre as técnicas existentes**.**

### 2.1 Base de Conhecimento

Em *Knowledge-Base Methods for WSD* [Agirre e Edmonds 2007 - Capítulo 5], Rada Mihalcea apresenta as pesquisas atuais no tocante ao uso de bases de conhecimentos, incluindo métodos que utilizam dicionários de definições, medidas de similaridades em redes semânticas, heurísticas e métodos baseados em propriedades da linguagem humana.

### 2.2 Não-supervisionada

Em *Unsupervised Corpus-Based Methods for WSD* [Agirre e Edmonds 2007 - Capítulo 6], Ted Pedersen explora métodos que não utilizam fontes externas (dicionários, por exemplo) para avaliar a similaridade entre as palavras. Essa abordagem determina os significados das palavras com base no contexto em que as palavras ocorrem. Isto é baseado na hipótese de que, se as palavras são utilizadas em contextos semelhantes terão significados semelhantes. O autor, também, apresenta duas abordagens distributivas com corpora monolíngues e métodos baseados em equivalência de translação, com o uso de corpora paralelo.

Em *Automatic Acquisition of Lexical Information and Examples* [Agirre e Edmonds 2007 - Capítulo 9], Julio Gonzalo e Felisa Verdejo consideram a aquisição de conhecimento como o gargalo dos métodos supervisionados. Segundo os autores, por tratar-se de uma área ainda pouco explorada, existe um leque de opção de pesquisas a ser realizada com essa abordagem.

### 2.3 Supervisionada

*Em Supervised Corpus-Based Methods for WSD* [Agirre e Edmonds 2007 - Capítulo 7], Lluís Màrquez, Gerald Escudero, David Martínex e German Rigau apresentam métodos que induzem os modelos de classificação ou regras a partir de documentos marcados manualmente. Essa é a abordagem dominante na atualidade. Os autores apresentam técnicas

de aprendizado de máquinas com a utilização de algoritmos de Redes Bayseanas e Máquina de Vetor de Suporte.

### 2.4 Combinadas

Em *WordNet Based Word Sense Disambiguation* [Siemiński 2011], Andrzej Siemiński propõe a utilização da técnica Colônia de Formigas no processo de desambiguação do sentido das palavras. Trata-se de uma técnica que faz uso de probabilidade para resolver problemas computacionais relacionados à otimização de caminhos.

### 2.5 Classificação Automática de Documentos

Em *Automatic Document Classification Temporally Robust* [Salles *et al* 2010], são apresentados mecanismos para Classificação Automática de Documentos robustos à evolução temporal. Este trabalho fundamenta a importância em analisar a dinâmica da linguística e conhecimento, afirmando que estes fatores influenciam diretamente o processo de classificação de documentos. Duas abordagens vêm se desenvolvendo, que são:

a) Na Classificação Adaptativa de Documento [Cohen e Singer 1999], um conjunto de técnicas que visam contornar os problemas relacionados aos aspectos temporais, melhorando a efetividade e a precisão dos classificadores por meio de adaptações incrementais, levando em consideração contexto, modelos incrementais e eficiência computacional dos classificadores;

b) Tendência de Tópicos [Tsymbal 2004], considera que conceitos e interesses dos usuários se modificam ao longo do tempo e que essas mudanças podem tornar inconsistentes modelos de classificação construídos com dados antigos.

## 3. Resultado Esperado

A relevância deste trabalho está relacionada a acompanhar o processo de evolução de conhecimento, com base em associações linguísticas, para um processo contínuo de classificação de documentos com o objetivo de obter uma desambiguação mais aperfeiçoada do sentido das palavras.

A pesquisa será desenvolvida sobre a hipótese de que o sentido das palavras é contextual (temporal, regional etc.) e, por isso, o processo de classificação é algo continuo e que, de acordo com o conhecimento adquirido, influencia a desambiguação dos termos.

Diferentemente das abordagens vistas até o momento em trabalhos sobre a área, as bases de conhecimento são universais e estáticas e, por isso, não levam em consideração o processo natural da evolução do conhecimento e a dependência contextual do sentido das palavras. Sabendo disso, propomos, neste trabalho, a utilização de uma base dinâmica de conhecimento, possibilitando a aquisição de novos termos aos domínios de conhecimento, possibilitando melhorias no processo de classificação automática de documentos.

A conclusão deste trabalho apresentará uma nova forma de abordar o problema de classificação de documentos com vocabulários temáticos dinâmicos para a desambiguação de termos, tal que este processo irá considerar a relatividade do conhecimento humano. Sendo assim, a abordagem se mostrará aplicável em domínios abrangentes e múltiplos contextos,

fazendo com que se torne uma solução flexível e ajustável às necessidades de qualquer problema ao qual se deseje fazer uso.

## Referencias


Agirre, E., Edmonds, P. (2007) "Word Sense Disambiguation: Algorithms and Applications", Springer, New York, NY.

Cohen, W. W., Singer, Y. (1999) "Context-sensitive learning methods for text categorization". ACM Trans. Inf. Syst., 17(2).

Gliozzo, A.; B. Magnini; C. Strapparava. (2004) "Unsupervised domain relevance estimation for word sense disambiguation". In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. EMNLP. Barcelona, Spain.

Mallery, J. C. (1988) "Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers". Ph.D. dissertation. MIT Political Science Department, Cambridge, MA

Navigli, R. (2009) "Word Sense Disambiguation: A Survey". ACM Computing Surveys.

Nunes, J. O. (2009) "Comunicação, sociedade e novas tecnologias: bases de transformação para novas práticas de produção e recepção do jornalismo on-line". In: Fuser, B.; Pernisa, C. (Org.). Comunicação e Tecnologias. 1. E-papers, Rio de Janeiro, v. 1, p. 115-125.

Russel, S. , Norvig, P. (2003) "Artificial Intelligence: A Moderrn Approach". Prentice Hall, 2th edition.

Salles, T., Rocha, L., Mourão, F., Pappa, G. L., Cunha, L., Gonçalves, M. A., Meira Jr, W. (2010) "Automatic Document Classification Temporally Robust". Journal of Information and Data Management, Vol. 1, No. 2, Pages 199–211.

Siemiński, A. (2011) "WordNet Based Word Sense Disambiguation". ICCCI'11 Proceedings of the Third International Conference on Computational Collective Intelligence: Technologies and Applications.

Silva, W. J. da., Baptista, C. S. Schiel, U., Silva, E. R. da, Menezes, L. C. de. Fernandes, R. M. (2006) "Freebie: Uma Biblioteca Digital Baseada em Software Livre com Suporte a Buscas Textual e Espacial". In: WebMedia 2006, Natal.

Tristão, A. M. D., Fachin, G. R. B., Alarcon, O. E. (2004) "Sistema de classificação facetada e tesauros: instrumentos para organização do conhecimento". Ci. Inf., Brasília, v. 33, n. 2, p. 161-171.

Tsymbal, A. (2004) "The problem of concept drift: Definitions and related work". Technical Report TCD-CS-2004-15, Computer Science Department, Trinity College, Dublin