UMA ABORDAGEM ONTOLÓGICA PARA A INDEXAÇÃO DE DOCUMENTOS ELETRÔNICOS

Cláudio Gottschalg-DUQUE[1] (UnB)

RESUMO: Este trabalho trata da indexação automática de documentos eletrônicos disponibilizados em língua portuguesa. Apresenta-se uma proposta de desenvolvimento e avaliação de um Sistema de Processamento de Linguagem Natural cuja aplicação envolve a análise sintática e semântica de documentos. A indexação é baseada na extração de etiquetas sintáticas das palavras que compõem os documentos para a geração de etiquetas semânticas dessas palavras, para então gerar automaticamente uma ontologia leve. Um protótipo e um *corpus* foram utilizados visando atestar que é possível desenvolver e implementar um Sistema de Recuperação de Informação totalmente baseado em teorias lingüísticas, teorias de lingüística computacional e ontologia.

ABSTRACT: This paper is about automatic indexation of electronic documents in Portuguese language. A development and evaluation of a Natural Language Processing System, whose application involves the syntactic and semantic analysis of documents, is proposed. The indexation is based on the extraction of syntactic labels of the words that compose the documents for the generation of semantic labels of those words, for then to generate a lightweight ontology automatically. A prototype and a corpus were used seeking to attest that it is possible to develop and to implement an Information Retrieval System totally based on linguistic theories, theories of computer linguistics and ontology.

1. Introdução

A área de Recuperação de Informação (RI) cresceu enormemente em importância, particularmente devido ao aumento da disponibilidade de informação em formato digital (RIJSBERGEN, 1979; WITTEN, 1999; GARFIELD, 2001). As publicações eletrônicas, por exemplo, merecem uma atenção especial em relação ao meio em que são disponibilizadas (GOTTSCHALG-DUQUE, 1998; ROSENFELD & MORVILLE, 2002) e os documentos digitalizados são um desafio para os especialistas em sistemas de recuperação de informação por, na maioria das vezes, tratarem-se de imagens, requerendo sistemas especialistas capazes de processar imagens digitais associadas às informações textuais (DE ANDRADE & ARAUJO, 2000). A coleção de documentos encontrada em bancos de dados pode ser indexada e recuperada de várias maneiras, dentre as quais através da utilização de técnicas inteligentes que permitam o mapeamento automático das palavras por categorias (HONKELA et al., 1996).

As Bibliotecas Digitais, que são repositórios digitais de todos os tipos de informação, variando de materiais históricos convertidos, como textos manuscritos digitalizados, a tipos de informação que não tem nenhum análogo no mundo físico, são de algum modo, muito diferentes das Bibliotecas Tradicionais. Contudo, elas são notavelmente semelhantes, pois as pessoas ainda criam informação, basicamente veiculadas através de texto, que tem que ser organizada, armazenada e distribuída, e elas ainda precisam encontrar e utilizar a informação que outros criaram (NÜRNBERG et ali, 1995; ARMS, 2000).

A World Wide Web evolui semanticamente (BERNERS-LEE, HENDLER and LASSILA, 2001), ela é o catalisador de todas as mídias, sugerindo a discussão sobre a validade ou não das abordagens cartesianas a novos problemas relacionados à informação, é o próprio "Ponto de Mutação" (CAPRA, 2001). É o meio que pode efetivamente conduzir a sociedade da informação a superação das desigualdades sociais e a agregação de valores econômico-político-culturais (TAKAHASHI, 2000). Entretanto, simultaneamente, ela indica um impasse perigoso e sugere boas perspectivas para o futuro, possibilitando uma nova visão da realidade, que envolve mudanças radicais em nossos pensamentos, percepções e valores.

Este projeto releva todas as questões citadas anteriormente e apresenta o desenvolvimento de novas técnicas de indexação baseadas em técnicas utilizadas em diferentes áreas, objetivando aperfeiçoar a qualidade da informação recuperada. Recuperação de Informação é entendida como o clássico problema da recuperação efetiva e eficiente de documentos pertinentes extraídos de uma grande coleção de acordo com uma específica necessidade de informação de um usuário (ROWLEY, 1996; BAEZA-YATES & RIBEIRO-

[1] klaussherzog@gmail.com, ou, klauss@unb.br

NETO, 1999 WITTEN et ali, 1999). Neste trabalho focalizamos em informação textual, apesar da existência e ampla divulgação de outros formatos de informação digital como, por exemplo, imagens.

A Recuperação de Informação na Web consiste de três processos: coleta, indexação e ordenação (GUDIVADA et ali, 1997; LAWRENCE, GILES & BOLLOCKER, 1999; CLEVELAND & CLEVELAND, 2000; LOSADA & BARREIRO, 2000; GLOVER et ali, 2002; LAENDER et ali, 2002). Enquanto que o processo de coleta está basicamente resolvido os outros dois processos não estão. A indexação é um processo que consiste em nomear palavras-chave de um documento (as palavras-chave são a representação do documento) e o processo de ordenação consiste em disponibilizar os documentos de acordo com uma graduação que condiz com as representações que satisfaçam as necessidades do usuário.

Um SRI, apesar da evolução dos sistemas de hardware, não pode incorporar todo o conteúdo da informação de uma coleção de documentos, porque isto compromete a eficiência do sistema, que trata de processar a representação do conteúdo informacional da coleção. Indexar normalmente é a determinação da representação do documento. A conversão da consulta do usuário para uma representação também é possível, permitindo assim o processo de ordenação dos documentos contidos na coleção em função da consulta do usuário. Assim, a forma e a maneira de representar a informação para a devida disponibilização é fundamental para todo o SRI.

Propõe-se o desenvolvimento, implementação e avaliação de um Sistema de Processamento de Linguagem Natural que seja robusto, um protótipo que execute análise sintática e semântica de textos de Ciência da Informação escritos em português do Brasil, otimizando a indexação e ordenação dos mesmos.

Na 2 seção deste artigo é abordada a composição do sistema de recuperação de informação. A segunda seção apresenta o módulo de processamento de linguagem natural, seguido do módulo de geração de ontologia. A quarta seção é o módulo de geração de índice e a última seção aborda a discussão da proposta.

2. Composição do Sistema de Recuperação de Informação

O SRI é composto de três módulos, o Módulo de Processamento de Linguagem Natural (MPLN), o Módulo Gerador de Ontologias (MGO) e o Módulo Gerador de Índices (MGI) (veja figura 1). Durante as primeiras etapas de todo o processo, os textos são convertidos para o formato TXT. A seguir, são reconhecidas unidades lexicais que são etiquetadas com a informação apropriada (substantivo, determinante, verbo etc). Na próxima etapa, são identificadas as estruturas sintáticas e uma análise superficial é processada, indicando os sintagmas nominais e verbais, por exemplo. Após a identificação das estruturas sintáticas, são identificados os elementos semânticos e uma análise visando à identificação das relações semânticas (JACKENDOFF, 1990), é processada. Finalmente, uma “ontologia leve” para a coleção que está sendo indexada é criada e enviada para o módulo indexador. No módulo indexador, a “ontologia leve”, é analisada, o índice da coleção é criado e o processo é finalizado. O MPLN visa criar índices baseados em estruturas conceituais no lugar de índices baseados em termos de documentos. Isto significa que o MPLN processa o documento, extraindo descrições do texto através de seus sub-módulos. SMA: Sub-Módulo Atomizador. SMOSi: Sub-Módulo Sintático. SMOSe: Sub-Módulo Semântico. Os conceitos utilizados para a criação do índice dos documentos são pré-definidos pelo MGO: Módulo Gerador de Ontologias, que é formado pelos sub-módulos SMOB: Sub-Módulo de Ontologia Básica e SMOF: Sub-Módulo de Ontologia Formada. O MGO fornece subsídios para o MGI: Módulo Gerador de Índice que é composto dos sub-módulos SMRI: Sub-Módulo de Regras de Índice. SMEI: Sub-Módulo de Estrutura de Índice.

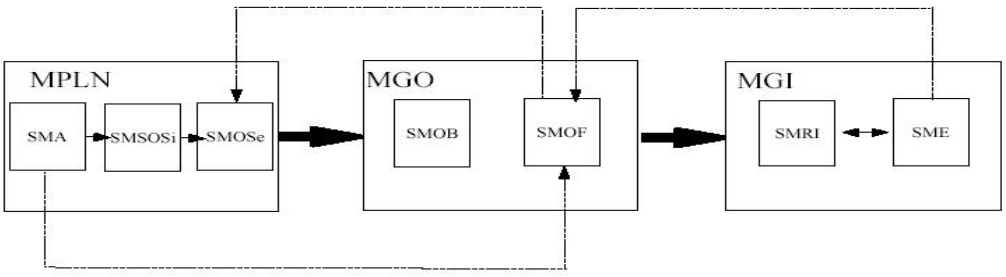

Figura 1. Sistema de Recuperação de Informação e seus respectivos módulos.

2.1 Módulo de Processamento de Linguagem Natural

A abordagem aqui discutida adota o Processamento de Linguagem Natural (PLN). Embora o PLN abranja várias e complexas áreas do conhecimento humano, como por exemplo lingüística e filosofia entre outras, (WINOGRAD, 1972; OKSEFJELL & SANTOS, 1998; NUNES et ali, 1999; RANCHHOD, 2001; MITKOV, 2003) neste trabalho será enfatizada a lingüística, objetivando a indexação automática dos documentos. A meta é aperfeiçoar o processo de indexação produzindo um índice de conceitos estruturados. O módulo de PLN deve analisar as orações nos documentos sintática e semanticamente de tal modo que permita facilmente a produção de índices de conceitos, resultando em uma ontologia básica. Um sistema de PLN deve ser robusto, para toda e qualquer oração processada *(input)*, uma resposta *(output)* deve ser produzida e em tempo hábil. Somente assim a adoção deste sistema fará efetivamente um diferencial positivo em relação às abordagens estatísticas usadas em SRI tradicionais (PERSIN, 1994; PÔSSAS et ali, 2002). Hoje em dia cada vez mais pesquisas indicam ou sugerem que o uso adequado de sistemas de PLN ajuda a diminuir significantemente os custos de indexação (KURAMOTO, 1995; SMEATON, 1997; OLTMANS, 1999; HIEMSTRA, 2001; PAULO et ali, 2002, REHM, 2002).

```
1-Para cada palavra i lida do texto
      Etiquetar palavra
2-Para cada sentenca i lida do texto
      Etiquetar sentenca
3- Repetir processo
      Enquanto uma palavra i lida do texto estiver sem etiqueta
      Enquanto uma sentenca i lida do texto estiver sem etiqueta
4-Extrair sentenca_autor sentença_titulo sentença_palavras-chave
5-Armazenar sentenca_etiquetada_SMA i em SMA_saida[i]
6-Armazenar sentenca_autor sentença_titulo sentença_palavras-chave em SMOF[i]
```

Figura 2. Algoritmo proposto para o módulo SMA.

```
1-Para cada palavra_SMA i lida de SMA_saida[i]
      Etiquetar palavra_SMA
2-Para cada sentença_SMA i lida de SMA_saida[i]
      Etiquetar sentença_SMA
3-Repetir processo
      Enquanto uma palavra_SMA i lida do texto estiver sem etiqueta
      Enquanto uma sentença_SMA i lida do texto estiver sem etiqueta
4-Armazenar sentenca_etiquetada_SMOSi i em SMOSi_saida[i]
```

Figura 3. Algoritmo proposto para o módulo SMOSi.

```
1-Para cada palavra_SMOSi i lida de SMOSi_saida[i]
      Etiquetar palavra_SMOSi
2-Para cada sentença_SMOSi i lida de SMOSi_saida[i]
      Etiquetar sentença_SMOSi
3-Repetir processo
      Enquanto uma palavra_SMOSi i lida do texto estiver sem etiqueta
      Enquanto uma sentença_SMOSi i lida do texto estiver sem etiqueta
4-Armazenar sentenca_etiquetada_SMOSe i em SMOSe_saida[i]
```

Figura.4. Algoritmo proposto para o módulo SMOSe.

2.2 Módulo de Geração de Ontologia

Ontologia é a especificação de uma conceitualização, que é uma visão abstrata e simplificada do universo que se pretende representar (GRUBER, 1993). A ontologia fornece um vocabulário comum de uma área e define, com diferentes níveis de formalismo, o significado dos termos e dos relacionamentos entre os mesmos (GOMEZ-PÉREZ & BENJAMINS; 1999). As ontologias são estruturadas de tal maneira que permitem um considerável ganho de qualidade quando empregadas num sistema de classificação além das possibilidades oferecidas por outros sistemas como *thesauri*. A idéia é que o índice criado a partir de estruturas conceituais geradas por meio do resultado de extensa análise de linguagem natural apresente um melhor desempenho para as respostas às consultas de usuários. Para a dedução, identificação e conseqüente extração destes conceitos são usadas técnicas de bases de conhecimento (FREDERIKSEN, 1975), que também permitem inferir a correlação dos conceitos extraídos.

```
1-Para cada conceito i lido de SMOSe_saida[i]
      Etiquetar conceito
2-Repetir processo
      Enquanto um conceito i lido do texto estiver sem etiqueta
3-Armazenar conceito i em SMOF_saida[i]
```

Figura 5. Algoritmo proposto para o módulo SMOF.

2.3 Módulo Gerador de Indexação

Durante as primeiras etapas de todo o processo, os textos são convertidos do formato no qual eles estão (normalmente PDF, PS, HTML ou DOC) em um formato canônico no qual são usadas etiquetas de XML (W3C, 1998; LOBIN, 2003) para delimitar e identificar as várias partes constituintes dos mesmos. A seguir são reconhecidas unidades lexicais que são etiquetadas com a informação apropriada (substantivo, determinante, verbo, etc). Na próxima etapa são identificadas as estruturas sintáticas e uma análise superficial é processada. Após a identificação das estruturas sintáticas são identificados os elementos semânticos e uma análise profunda é processada. Finalmente uma ontologia é criada ou atualizada e o processo de indexando é finalizado.

A meta do MPLN foi determinada em termos das restrições que a ontologia provê, fornecendo uma especificação de produção que consiste na informação lingüística extraída do texto.

Baseada nos termos extraídos dos textos e nas possíveis relações entre os mesmos foi criada uma ontologia de conceitos utilizados na Ciência da Informação. A organização do Sistema é modular, assim a parte responsável pelo Processamento de Linguagem Natural (Módulo de Processamento de Linguagem Natural) está estritamente separada da parte responsável pela geração do índice (Módulo Gerador de Índice, o módulo que tem acesso à ontologia). Isto significa que todo o processo lingüístico que é necessário para a criação da ontologia é realizado separadamente. Teoricamente, o conhecimento disponível na base de conhecimento é suficiente para produzir uma análise clara e imune a ambigüidades, caso contrário, o MPLN deve ser capaz de retornar como resposta blocos de frases que são consideradas possíveis opções que podem ser usadas pelo MGO e conseqüentemente pelo MGI.

```
1-Para cada conceito i armazenado em SMOF_saida[i]
      Identificar proposicoes i
      Identificar proposicoes i com núcleo ser ou estar
      Identificar objetos i das proposicoes i com núcleo ser ou estar
2-Etiquetar
      proposicoes_serouestar i
      objetos i
      proposicoes i
      agentes i
      instrumentos i
3-Criar lista invertida
      de proposicoes_serouestar i
      de proposicoes i
4-Armazenar lista invertida de proposicoes i e de proposicoes_serouestar i em
SMEI
```

Figura 3.6. Algoritmo proposto para o módulo SMRI.

3. Discussão

Apresentaram-se as características básicas de um Sistema de Recuperação de Informação que usa Processamento de Linguagem Natural e Ontologias. São usados *parsers* (analisadores) Sintático e Semântico que fornecem subsídios para permitir a criação de uma ontologia. As ontologias podem ser entendidas como ferramentas que permitem a construção de Sistemas de Bases de Conhecimento. Neste projeto uma ontologia é utilizada para a construção de representações expressivas de relações múltiplas de estruturas conceituais do texto e entre textos, com o intuito de facilitar a resposta às questões de consultas e de navegação em um SRI. São propostas contribuições para *parsers* sintáticos e aplicações de *parsers* semânticos. Acredita-se que a anotação (etiquetagem) do contexto sintático de textos melhora a identificação do relacionamento semântico existente entre as palavras, isto permite a identificação e criação de bancos de conhecimento, o que conjuntamente com as características inerentes das ontologias permitirá o desenvolvimento de um SRI mais rápido e eficiente.

4. Referências bibliográficas

ARMS, W. Y. Digital Libraries (Digital Libraries and Electronic Publishing). The Mit Press, Cambridge, MA, London, England, 2000.

BAEZA-YATES, R. & RIBEIRO;NETO, B. Modern Information Retrieval. Addison-Wesley Pub Co; 1st edition (May 1999).

BERNERS-LEE, T., HENDLER, J. & LASSILA, O. The semantic web. Scientific American, May 2001.

CAPRA, F. O. Ponto de Mutação. Editora Cultrix. 22 Edição. 2001.

CLEVELAND, D. B. and CLEVELAND, A.D. Introduction to Indexing and Abstracting. Libraries Unlimited, 3 edição, 2000.

DE ANDRADE, N.S., & ARAÚJO, A. de A. Multimídia para acesso a acervos históricos, Revista iP-Informática Pública, PRODABEL, Belo Horizonte-MG, Brazil, vol. 2, no. 1, 2000, pp 49-66.

FREDERIKSEN, C.. Representing Logical and Semantic Structure of Knowledge Aquired from Discourse. Cognitive Psychology 7, 1975, pp 371-458.

GARFIELD, E. A Retrospective and Prospective View of Information Retrieval and Artificial Intelligence in the 21st Century. Journal of The American Society for Information Science and Technology. 52(1), 2001.

GLOVER, E. J.; TSIOUTSIOULIKLIS, K.; LAWRENCE, S.; PENNOCK, D.M.; FLAKE,

G.W. Using Web Structure for Classifying and Describing Web Pages. Proceedings of WWW-02, International Conference on the World Wide Web. 2002.

GOMEZ-PÉREZ, A. & BENJAMINS, V.R. Overview of Knowledge sharing and reuse components: Ontologies and problem-solving methods. In: International Joint Conference on Artificial Intelligence (IJCAI-99), Workshop on Ontologies and Problem-Solving Methods (KRR5), V.R. Benjamins, et al., Editors. Stockolm,1999.

GOTTSCHALG-DUQUE, C. G. A Leitura em Ambiente Multimídia: A Produção de Inferências por parte do Leitor a partir da Compreensão de Hipertextos, FALE-UFMG em 16 de novembro de 1998.

GRUBER; T. R. A Translation approach to portable ontology specifications. Knowledge Acquisition, 1993.

GUDIVADA, V. N.; RAGHAVAN, V. V.; GROSBY, W. I.; KASANAGOTTU, R. Information Retrieval on the World Wide Web. IEEE Internet Computer, 1997.

HIEMSTRA, D. Using Language Models for Information Retrieval. Phd Thesis University of Twente,

Enschede, 2001.

HONKELA et al. Exploration of Full-Text Databases with Self-Organizing Maps. Proceedings of the ICNN96, International Conference on Neural Networks. 1996.

JACKENDOFF, R. Semantic Structures. Cambridge: MIT Press, 1990.

LAENDER, A. H. F., RIBEIRO-NETO, B., DA SILVA, A. S. and TEIXEIRA, J S. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record*, 2002.

LAWRENCE, S.; GILES, C. L. and BOLLOCKER, K. Digital Libraries and Autonomous Citation Indexing. IEEE Computer, Volume 32, Number 6, pp. 67-71, 1999.

LOBIN, H. Textauszeichnung und Dokumentgrammatiken. In: Texttechnologie.LOBIN, H & LEMNITZER, L. (ed.). Stauffenburg Verlag, 2003

LOSADA, D. E. & BARREIRO, A. Efficient algorithms for ranking documents represented as DNF formulas. In Proc. ACM SIGIR-2000 Workshop on Mathematical and Formal Methods in Information Retrieval. Pgs 16-24. Athens, Greece, July 2000.

MITKOV, R (ed.). The Oxford Handbook of Computational Linguistics. Oxford University Press; (March 2003).

NUNES, M. G. V.; DA SILVA, B. C. D.; RINO, L. H. M. ; OLIVEIRA JR, O. N.; MARTINS R. T.; MONTILHA, G. Introdução ao Processamento de Linguagens Naturais. Notas Didáticas do ICMC, São Carlos, 1999.

NÜRNBERG, P. J.; FURUTA, R.; LEGGETT, J. J.; MARSHALL, C. C.; SHIPMAN III, F. M. Digital Libraries: Issues and Architectures. Digital Libraries 95. 1995.

OKSEFJELL, S. & SANTOS D. Breve panorâmica dos recursos de português mencionados na Web. In Anais do III Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'98). Porto Alegre, 3-4 novembro 1998. pp.38-47.

OLTMANS, J. A. E. A Knowledge-Based Approach to Robust Parsing. PhD Thesis.Centre for Telematics and Information Technology (CTIT) P.O. Box 217, 7500 AE Enschede, The Netherlands, 1999.

PAULO, J. L.; CORREIA, M.; MAMEDE, N. J.; HAGÈGE, C. Using Morphological, Syntactical and Statistical Information for Automatic Term Acquisition. In: Advances in Natural Language Processing. Third International Conference, Proceedings, PorTAL 2002, Faro Portugal, June 23-26, 2002.

PERSIN, M. Document filtering for fast ranking. Proc. ACM SIGIR, Conf. Dublin, Ireland, 1994.

PÔSSAS, B.; ZIVIANI, N.; MEIRA, W.; RIBEIRO-NETO, B. Set-Based Model: A New Approach for Information Retrieval Proc. 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02), Tampere, Finland, August 2002.

RANCHHOD, E. (ed.). Tratamento das Línguas por Computador. Uma introdução à lingüística computacional e suas aplicações, Lisboa: Caminho, 2001.

REHM, G. Towards Automatic Web Genre Identification -- A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. Proceedings of the Hawai'i International Conference on System Sciences, January 7-10, 2002, Big Island, Hawaii.RIJSBERGEN, C. J. van. Information Retrieval. 1979. Disponível em “http://www.dcs.gla.ac.uk/Keith/Preface.html”

ROSENFELD, L. & MORVILLE, P. Information Architecture for the World Wide Web: Designing Large-Scale Web Sites. O'Reilly & Associates; 2nd edition (August 15, 2002).

SMEATON, A. F. Information Retrieval: Still Butting Heads with Natural Language Processing Springer-Verlag, Information Extraction - A multidisciplinary approach to an emerging information technology.
TAKAHASHI, T. (org.). Sociedade da Infomação no Brasil Livro Verde. Brasília Ministério da Ciência e Tecnologia, 2000.

WINOGRAD, T. Understanding Natural Language, (191 pp.) New York: Academic Press, 1972.

WITTEN; I. et al. Managing Gigabytes. Morgan Kaufmann Publishers, Inc. Second Edition, 1999.