Parlatorium
revista eletrônica da FAMINAS-BH

MEDIDAS DE SIMILARIDADE EM DOCUMENTOS ELETRÔNICOS

Luiz Cláudio Gomes Maia[1]
Renato Rocha Souza[2]

RESUMO: Algoritmos e técnicas aplicáveis na Recuperação da Informação na forma eletrônica estão em evolução e representam uma grande fatia dos estudos recentes em Ciência da Informação em conjunto com outras áreas como a Ciência da Computação. A Web, através de sua estrutura não linear formada por hiperlinks, ampliou as possibilidades, anteriormente limitada ao texto, de resultados mais satisfatórios com o uso da análise de ligações. Este artigo faz uma compilação de técnicas de Recuperação de Informação e medidas de similaridade em um conjunto de documentos eletrônicos.

Palavras-chave: agrupamento automático de documentos; algoritmos de treinamento; similaridade.

ABSTRACT: Applicable algorithms and techniques in the Information Retrieval in the electronic form are in evolution and represent a great slice of the recent studies in Information Science in set with other areas as the Computer science. The Web, through its not linear structure formed of hyperlinks, extended the possibilities, previously limited to the text, of more satisfactory results with the use of the analysis of links. The bibliometric is an example of the use of the analysis of links. The research presents measured experiments involving of similarity in a set of electronic documents.

Keywords: *bibliometric; link analysis; similarity of electronic documents.*

[1] Professor da Faculdade de Tecnologia INED, mestre e doutorando do programa de Ciência da Informação da UFMG.
[2] Professor da Escola de Ciência da Informação da UFMG. Doutor em Ciência da Informação pela UFMG.

1 INTRODUÇÃO

A informação cada vez mais é registrada diretamente em meios digitais. Vivencia-se uma consolidação, não só da convergência digital, mas também da criação de conteúdos já digitalizados. Neste contexto, a publicação e criação de conteúdos tornam-se mais fáceis e, por consequente, informações irrelevantes, de baixa qualidade e mesmo de baixa confiabilidade fazem parte de um "lixo informacional" crescente e que preocupa toda a sociedade.

Um dos principais campos de estudo da Ciência da Informação compreende o tratamento e organização da informação de forma a possibilitar resultados de busca satisfatórios, atendendo a demanda do usuário, sem a interferência do "lixo informacional".

Um Sistema de Recuperação da Informação (SRI) deve analisar os documentos para saber os itens de seu acervo que são relevantes frente a uma consulta do usuário. O objetivo é atender de forma satisfatória ao usuário. Para isto, pesquisas envolvendo técnicas e algoritmos aplicáveis em SRI são constantes. Atualmente, com todo o aporte computacional disponível, programas de computador podem se valer de um processamento rápido para melhorar ainda mais a satisfação do usuário no uso destes sistemas.

Algoritmos envolvendo métricas informacionais aplicáveis em SRI estão em evolução e representam uma grande fatia dos estudos recentes em Ciência da Informação, em conjunto com outras áreas, como a Ciência da Computação.

Este artigo realiza uma compilação de técnicas atuais de Recuperação da Informação, e propõe medidas para realização de agrupamento por similaridade (*clustering*) e classificação de documentos eletrônicos.

Estas medidas permitem uma análise automatizada da similaridade de documentos eletrônicos, o que pode redundar em projetos inovadores de sistemas de recuperação de informação.

2 ANÁLISE DE TEXTO

A análise de texto (*text analysis*) corresponde a uma área que envolve outras subáreas como, por exemplo, a mineração de texto (*text mining)* e a área de Processamento de Linguagem Natural (PLN). A PLN também é uma subárea da inteligência artificial e da linguística que estuda os problemas da geração e tratamento automático de línguas humanas naturais.

A Mineração de texto (*Text Mining*) refere-se ao processo de obtenção de informação a partir de texto em línguas naturais. Se praticada em conjunto com a mineração de dados, que consiste em extrair informação de bancos de dados estruturados, a mineração de texto extrai informação de dados não estruturados ou semi-estruturados.

O texto corresponde à principal parte das muitas que podem compor um documento, e seu tratamento, como um processo de criação dos índices, é explorado pelos SRIs.

3 CONSTRUÇÃO E ARMAZENAMENTO DO ÍNDICE

O índice tem como objetivo a recuperação rápida da informação. A forma como se constrói, armazena e manipula o índice muda de acordo com a tecnologia empregada e por

consequente sua evolução. Tradicionalmente, CPUs eram lentas e a utilização de técnicas de compactação não seria interessante. Hoje, as CPUs já são mais rápidas, entretanto temos um armazenamento em disco rígido lento, que, para ser contornado, necessitamos diminuir o espaço de armazenamento ou mesmo utilizar memórias mais rápidas (na hierarquia) como a RAM.

Basicamente, a criação do índice significa criar um dicionário de palavras utilizadas em todos os documentos da coleção e criar um índice invertido indicando em qual documento cada palavra aparece.

Com a criação deste índice torna-se extremamente mais rápido a busca de informações do que recorrer a varrer todos os textos palavra por palavra.

A maior parte dos SRI tem como base o modelo clássico ou o modelo estruturado.

Nos modelos clássicos, cada documento é descrito por um conjunto de palavras-chave representativas, também chamadas de termos de indexação, que buscam representar o assunto do documento e sumarizar seu conteúdo de forma significativa. (BAEZA-YATES; RIBEIRO-NETO, 1999).

Nos modelos estruturados, podem-se especificar, além das palavras-chave, algumas informações acerca da estrutura do texto. Estas informações podem ser as seções a serem pesquisadas, fontes de letras, proximidade das palavras, entre outras.

Dentre os modelos clássicos, temos o booleano, o vetorial e o probabilístico. O modelo booleano é baseado na teoria dos conjuntos e possui consultas especificadas com termos e expressões booleanas. Nas consultas são utilizados operadores lógicos como E, OU, NÃO para filtragem do resultado.

Apesar de ser um modelo bastante simples e muito utilizado ele apresenta as seguintes desvantagens, segundo Baeza-Yates e Ribeiro (1999):

- A recuperação é baseada numa decisão binária sem noção de casamento parcial;
- Nenhuma ordenação de documentos é fornecida;
- A passagem da necessidade de informação do usuário à expressão booleana é considerada complicada;
- As consultas booleanas formuladas pelos usuários são frequentemente simplistas;
- Em conseqüência o modelo booleano retorna poucos documentos em resposta às consultas;
- O uso de pesos binários é limitante.

Para contornar estas limitações, novos modelos são desenvolvidos tendo como base alguns destes modelos clássicos.

O modelo que permite localizar similaridade entre documentos é o vetorial. O vetor é definido através do conjunto de documentos que formam o corpora.

Todo o texto dos documentos é extraído e convertido em um formato que permita a fácil manipulação. Toda ordem das palavras é ignorada, o que pode ser interpretado como colocar todas as palavras de cada documento em um saco separado (a expressão *bag of words*). Todas as palavras em cada saco são contadas (processo de indexação) e o número de vezes que cada palavra aparece (forma mais simplista de dar valor ao peso) é armazenado em um vetor termo-por-documento.

Ele é arranjado de forma que cada linha representa uma palavra (termo) e cada coluna representa um documento. Os valores contem o peso dos termos para cada documento. Em geral, este tipo de vetor é extenso e a maioria dos pesos dos termos é zero.

	d1	*D2*	*d3*	*d4*	*d5*	*d6*	*d7*	*D8*	*d9*
Rede	*0*	*0,60*	*0*	*0,20*	*0,75*	*0,02*	*0*	*0,15*	*0,80*
Social	*0,20*	*0*	*0,05*	*0,30*	*0,75*	*0*	*0,02*	*0*	*0*
Pesquisa	*0*	*0,40*	*0*	*0,50*	*0*	*0*	*0*	*0*	*0,20*
Vetor	*0,20*	*0*	*0*	*0*	*0*	*0*	*0,10*	*0,10*	*0*

Tabela 1 – Exemplo do Modelo Vetorial

Nas colunas estão representados os pesos de cada termo no documento. No exemplo acima o termo Rede tem o peso de 0,75 no documento 5 enquanto que o termo "Pesquisa" não aparece no documento 3. Portanto, seu peso é 0.

Sobre o uso de pesos no modelo vetorial, Baeza-Yates e Ribeiro-Neto (1999) apresentam algumas considerações:

- Pesos não binários podem considerar mais adequadamente *matchings* parciais;
- Estes pesos são utilizados para calcular um grau de similaridade entre a consulta e o documento;
- A fórmula com que são calculados os pesos varia dentre as implementações;

Cada documento (coluna) pode ser considerado como um vetor ou uma coordenada em um espaço do vetor do multidimensional em que cada dimensão representa um termo.

O *term frequency* (TF) corresponde ao número de vezes que o termo aparece no documento. A equação é dada por:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Sendo:

$N_{i,j}$ é o numero de ocorrências do termo no documento J e o denominador corresponde ao número de ocorrências de todos os termos em J.

Já o IDF é uma medida de grande importância para complementar a equação acima já que avalia a importância do termo na coleção. É obtida dividindo a quantidade de documentos pelo número de documentos contendo o termo e então obtendo o logaritmo do resultado.

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

Sendo:

|D | é o total de documentos no corpus

$|\{d_j : t_i \in d_j\}|$ número de documentos onde o termo *ti* aparece.

Através da união das duas tem-se TF-IDF:

$$tfidf_{i,j} = tf_{i,j} . idf_i$$

A medida *term frequency–inverse document frequency* (TF-IDF) corresponde a uma medida estatística utilizada para avaliar o quanto uma palavra é importante para um documento em relação a uma coleção (*corpus*). Esta importância aumenta proporcionalmente com o número de vezes que a palavra aparece e diminui de acordo com o a frequência da palavra na coleção.

Dependendo da aplicação e experimento, a partir do modelo TF-IDF podem surgir outros modelos que modificam a sistemática de atribuição de pesos.

A análise semântica latente é uma técnica da PLN, relacionada a manipulação de vetores do índice. Ela está relacionada à aplicação da matemática para analisar a relação entre termos e documentos e decompor o vetor de índice. O processo matemático utilizado é o SVD (*Simple Value Decomposition*).

Alguns autores e pesquisas também a chamam de *Latent Semantic Indexing* (LSI).

A LSA trabalha com a sinonímia e a polissemia. Por exemplo, para a consulta "extravio de bagagem", feita a uma ferramenta de busca que usa LSA, o sistema retornará

documentos que contenham as frases "extravio de bagagem" e "extravio de mala", já que “bagagem” e “mala” têm o mesmo significado no contexto. Da mesma forma, em uma consulta por "banco de dados", o resultado da consulta incluirá somente documentos que contenham uma relação com "banco de dados", excluindo documentos que se referem à “banco” como móvel e “banco” como entidade financeira.

A LSA trabalha com vários vetores, criando desta forma uma matriz, que nas linhas estão representados os termos indexados de cada documento e nas colunas o documento. Desta forma é criada a relação à matriz termo-documento. Explicando melhor esta relação, seja *ti* a linha e *dj* a coluna da matriz, e seja o elemento da matriz *Oij* que representaria o número de vezes que o termo i aparece no documento j.

Após ser criada esta matriz termo-documento, é aplicado o *Simple Value Decomposition* – SVD. Essa decomposição divide a matriz termo-documento em três matrizes: a matriz U que contém os termos, a matriz S que contêm os valores mais representativos da matriz termo-documento (os valores singulares) e a matriz V que contém os documentos. Depois de criadas estas três matrizes, é escolhido um tamanho (nível k) para trabalhar com as mesmas. Escolhido este valor, são criadas três matrizes (que serão chamadas U', S' e V') de nível k, a estas três novas matrizes é multiplicado o vetor Q, que representa uma consulta. O resultado desta multiplicação será um vetor cujo conteúdo é uma lista dos documentos mais relevantes para a consulta fornecida.

De acordo com Ramsden (1974, p. 3), o termo “linguagens naturais” é comumente associado à linguagem falada e à linguagem escrita. É possível em indexação utilizar a linguagem natural simplesmente como é falada ou usada nos documentos sem tentar, por exemplo, controlar sinônimos ou indicar os relacionamentos entre os termos. Um índice feito desta maneira chama-se índice de linguagem natural. Como alternativa ao índice de

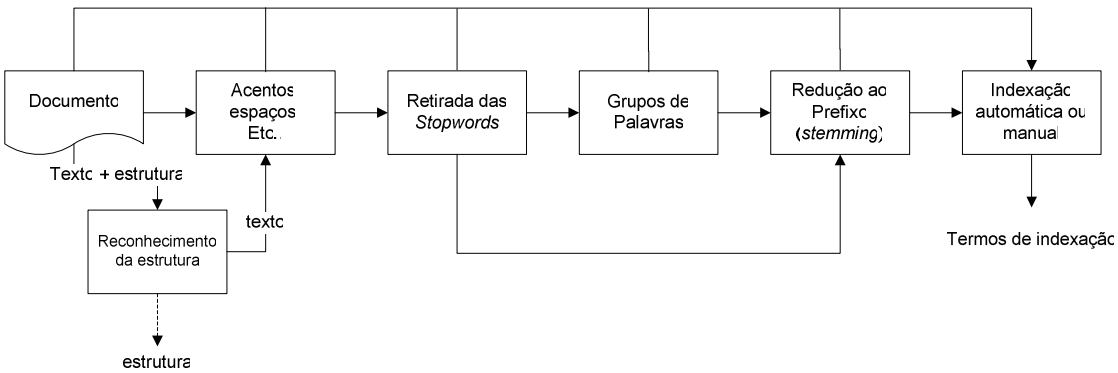
linguagem natural, pode-se usar uma linguagem artificial diante das necessidades do sistema de classificação, ou seja, uma linguagem de indexação. Em resumo,

> Esta linguagem refletirá um vocabulário controlado para o qual foram tomadas decisões cuidadosas sobre os termos a serem usados, o significado de cada um e os relacionamentos que apresentam (RAMSDEN, 1974, p. 3).

Existem contextos em que se pode utilizar uma linguagem de indexação – sistemas de classificação, listas de cabeçalhos de assunto, tesauros, etc. –, sendo que elas consistem de um vocabulário controlado e uma sintaxe a ser seguida.

O processo de indexação visa a representação dos conteúdos dos documentos, tendo como resultado uma lista de descritores. Ou seja, este processo tem como objetivo extrair os dados contidos nos documentos, organizando-as para permitir a recuperação destes últimos. Assim, os descritores devem ser, na maior extensão possível, portadores de informação, de maneira a relacionar um objeto da realidade extralinguística com o documento que contenha informações sobre este objeto. Contudo, na maioria dos SRI convencionais, os descritores representam com muita limitação as informações presentes no documento.

Alguns termos que podem prejudicar a recuperação, conhecidos como *stopwords*[3], são extraídos do texto através de um processo de tratamento do documento conforme ilustrado na figura abaixo.



[3] Palavras que não são úteis para recuperação de informações (e.g. palavras comuns, preposição, artigos, etc..)

Figura 1 – Fases do processamento do documento para submissão a indexação.
FONTE: Adaptado de BAEZA-YATES & RIBEIRO-NETO, 1999, p. 166.

Ao final do processamento têm-se, através de um processo de indexação automática ou manual, os termos de maior relevância para indexação. Técnicas como a de *Stemming* devem ser utilizadas para reduzir a redundância semântica entre os termos.

4 MEDIDAS DE SIMILARIDADE EM DOCUMENTOS ELETRÔNICOS

Os algoritmos que retornam similaridade entre documentos trabalham com métricas que retornam o quanto um documento é similar a outro. Existem diversos algoritmos e métricas utilizados em fins diversos. Um algoritmo deste tipo pode, por exemplo, ser utilizado na grade de programação digital da televisão para fornecer programas similares ao gosto do usuário, conforme demonstrado por Fabio (SANTOS SILVA, 2005) em projeto denominado Sistema de Recomendação Personalizada de Programas de TV (SRPTV).

No campo da estatística, temos duas medidas de similaridade básicas que se expandem para outros estudos: correlação e coseno. A correlação de Person (ou só correlação) entre dois vetores retorna um valor entre 0 e 1. Se for 1 eles estão fortemente correlacionados, isto é, os valores de um vetor podem predizer os valores do outro. Se for 0 não existe correlação. E se for -1 existe uma correlação inversamente proporcional.

O coseno é similar à correlação, retornando valores entre 0 e 1. Ele mede o ângulo entre dois vetores num espaço vetorial. Quanto mais próximo de 1 for o valor, mais similares são os dois vetores.

Para se localizar a similaridade entre dois documentos em um SRI utilizando VSM, calcula-se o coseno do ângulo formado no vetor termo-por-documento. No VSM padrão quanto menor o ângulo, mais próximo de 1 será o coseno e mais similar será o documento em relação a aquele termo.

$$sim(\vec{d_1}, \vec{d_2}) = \cos(\widehat{\vec{d_1}\vec{d_2}}) = \frac{\vec{d_1} \cdot \vec{d_2}}{|\vec{d_1}| \cdot |\vec{d_2}|} = \frac{\sum_i w_{i,1} \cdot w_{i,2}}{\sqrt{\sum_i w_{i,1}^2} \cdot \sqrt{\sum_i w_{i,2}^2}}$$

Sendo:

$w_{i,j}$ é o peso do temo t_i no documento d_j

Baeza-Yates e Ribeiro-Neto (1999) nos apresentam algumas outras observações sobre este modelo como um todo:

- Um conjunto ordenado de documentos é retornado, fornecendo uma melhor resposta à consulta.
- Documentos que têm mais termos em comum com a consulta tendem a ter maior similaridade;
- Aqueles termos com maiores pesos contribuem mais para o casamento do que os que têm menores pesos;
- Documentos maiores são favorecidos;
- A similaridade calculada não tem um limite superior definido.

O uso de um SRI e de um algoritmo de *clustering* para agrupar documentos envolve calcular a distância entre estes documentos na matriz. Existem, além do coseno de

similaridade, outras medidas, sendo que a distancia euclidiana é também muito utilizada. A distância euclidiana entre dois documentos d1 e d2 é definida por:

$$d(\vec{d_1}, \vec{d_2}) = \sqrt{\sum_i (w_{i,1} - w_{i,2})^2}$$

Sendo: $w_{i,j}$ é o peso do temo t_i no documento d_j.

A distância euclidiana necessita que quatro condições, nos vetores x, y e z, sejam validas para atuar como medida:

1. $d(x, y) \geq 0$

2. $d(x, x) = 0$

3. $d(x, y) = d(y, x)$

4. $d(x, z) \leq d(x, y) + d(y, z)$

Mais uma vez o tamanho do documento tem grande influência quando se utiliza a distância euclidiana.

O algoritmo de Rocchio (ROCCHIO, 1971 *apud* HARMAN, 1992) é um algoritmo de lote, que produz um novo vetor de pesos w a partir de um vetor de pesos existente w_1 e de um conjunto de exemplos de treinamento. O j^{th} componente w_j do novo vetor de componentes é (LEWIS, 1996):

$$w_j = \alpha w_{1,j} + \beta \frac{\sum_{i \in C} x_{i,j}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{i,j}}{n - n_C}$$

Sendo:

$$C = \{1 \leq i \leq n : y_i = 1\}$$

Onde n é o número de exemplos de treinamento, C é o conjunto de exemplos de treinamento positivos e é número de exemplos positivos de treinamento. Os parâmetros α, β e *Y* controlam o impacto relativo do vetor de pesos original, dos exemplos positivos e dos negativos respectivamente.

O algoritmo de Rocchio baseia-se na satisfação através do *feedback* do usuário com os resultados apresentados (treinamento positivo). Pode-se fazer uma relação com as técnicas de *Relevance Feedback* apresentadas e discutidas por Buckley (1995).

Para Buckley, *Relevance Feedback* é o processo automático de refinamento de uma consulta inicial, utilizando informações fornecidas pelo usuário sobre a relevância dos documentos previamente recuperados (em uma consulta anterior). Através do processo de retroalimentação, que corresponde a aplicar a equação apresentada, serão obtidas definições cada vez mais apuradas para as categorias envolvidas.

A medida kNN é definida por Yang em 1994 (*apud* CALADO *et al*., 2006) e definida por este nome na pesquisa de Calado (*et al.*, 2006) devido a se basear em testes realizados com categorias (k) vizinhas (*nearest neighbor*) e através de um processo de afunilamento definir a categoria.

A seguinte equação ilustra o algoritmo kNN:

$$S_{c_i,d} = \sum_{d' \in N_k(d)} similaridade\,(d,d') f(c_i,d')$$

Sendo:

K é igual ao número de vizinhos, *Nk*(*d*) corresponde aos documentos mais similiares a *k. e f*(*ci, d_*) corresponde a uma função binária que retorna se o documento *d'* pertence a uma categoria *ci* ou não.

O objetivo é filtrar os documentos baseado na predominância dos k vizinhos mais próximos. Os vizinhos mais próximos são os documentos que possuem maior valor de similaridade.

Algumas métricas utilizadas para identificação de dados similares são *Edge Cover, Shingsem, shingcom,* Distância de edição, *Similarity flooding, Shingles* e Série temporal.

Muitos algoritmos de agrupamento requerem como parâmetro predefinido o número de grupos, ou então outro parâmetro para definir a granularidade. A definição do número de grupos pode apresentar dificuldades de acordo com o conjunto de medidas e técnicas utilizadas. Existem alguns métodos e algoritmos para definir a quantidade de grupos de forma automática. Como, por exemplo: método baseado na distância, *dendrogram, Curvas de Sihouette, Bem-Hur, Elisseeff* e *Guyon*.

5 PESQUISAS SIMILARES

Esta pesquisa utiliza um modelo proposto por SOUZA (2005), em que o autor propõe o uso de sintagmas nominais como descritores para recuperação de documentos.

Calado (*et al.*, 2006) realiza um experimento utilizando as medidas de similaridade: *Amsler*, *Bibliographic Coupling*, Co-Citacion, kNN, SVM e *Naive Bayes* utilizando um corpora baseado no diretório de busca CADE. A pesquisa conclui que são necessárias novas experiências em outros corpos de documentos.

REFERÊNCIAS

ANDERSON, J.,; PEREZ-CARBALLO, J.. The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. **Part I: Research, and the Nature of Human Indexing. Information Processing and Management**, n. 37, 2001. p. 231-254.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval.** New York: ACM Press, 1999.

BUCKLEY, C.; SALTON, G.. Optimization of Relevance Feedback Weights In: **Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.** Washington: USA. 1995. p. 9-13.

CALADO, P. P.; CRISTO, M.; MOURA, E. S.; GONÇALVES, M. A.; ZIVIANI, N.; RIBEIRO-NETO, B.. Linkage similarity measures for the classification of Web documents. **Journal of the American Society for Information Science and Technology** (JASIST), vol. 57, no. 2, 2005. p. 208-221.

GREENBERG, J.. Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. **Journal of Internet Cataloging,** 6(4), 2004, p. 59-82.

HARMAN, D.. Relevance feedback and other query modification techniques. In William B. Frakes and Ricardo Baeza-Yates. **Information Retrieval:** Data Structures and Algorithms, p. 241-263. Prentice Hall, 1992.

IRVIN, K.K.. **Comparing Information Retrieval Effectiveness of Different Metadata.** Generation Methods. A Master's paper for the M.S. in I.S. degree. April, 2003.

JANSSENS, F.. Clustering of scientific fields by integrating text Mining and bibliometrics, **Katholieke Universiteit Leuven:** Faculteit Ingenieurswetenschappen. Mei, 2007.

KOCH, I. V.; SILVA, M.C.P.S. **Lingüística aplicada ao português:** sintaxe. São Paulo, Cortez, 1985.

KURAMOTO, H.. Sintagmas Nominais: uma nova proposta para a Recuperação da Informação. **DataGramaZero**, v. 3, n. 1, fev. 2002.

_____. Uma abordagem alternativa para o tratamento e a recuperação da informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, p. 182-192, v. 25, n. 2, maio/ago. 1996.

KWASNIK, B.H.. The role of classification in knowledge representation and Discovery. **Library Trends,** p. 22-47, v. 48, n. 1, Summer, 1999.

LAKOFF, G.. **Women, fire, and dangerous things:** what categories reveal about the mind. Chicago: The University of Chicago Press, 1987.

LAWRENCE, S.; GILES, C.. Accessibility of Information on the Web. **Nature,** p.107-109, n. 400, 1999.

MCCALLUM, A. K.; et al.. Automating de Construction of Internet Portals with Machine. **Learning Information Retrieval**, p. 127-163 , v.2, n. 3, 2000.

PERINE M.A.; et al.. O Sintagma Nominal em Português: Estrutura, Significado e Função, **Revista de Estudos da Linguagem**. n. esp.. 1996.

POMBO, O.. Da Classificação dos Seres à Classificação dos Saberes, Leituras. **Revista da Biblioteca Nacional de Lisboa,** n. 2, Primavera, pp. 19-33. disponível no site: http://www.educ.fc.ul.pt/docentes/opombo/investigacao/opombo-classificacao.pdf consultado em 05/12/2003

RAMSDEN, M. J.. **An introduction to index language construction, a programed text.** London: C. Bingley, 1974.

SALTON, G.. **Automatic information organization and retrieval.** New York: McGraw-Hill, 1968.

SANTOS SILVA, F.. Personalização de Conteúdo na TVDI através de um Sistema de Recomendação Personalizada de Programas de TV (SRPTV). **Anais...** III Fórum de Oportunidades em Televisão Digital Interativa, Poços de Caldas, 2005.

SATIJA, M.P.. Library classification:an essay in terminology. **Knowledge organization,** p. 221-229, v. 27,n. 4, 2000.

SOUZA, J.S.. **Classificação:** sistemas de classificação bibliográfica. 2.ed. São Paulo: Departamento Municipal de Cultura, 1950.

SOUZA, R.R.. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.** 2006. Tese (Doutorado) – Universidade Federal de Minas Gerais. Escola de Ciência da Informação.