Classificação Automática de Textos usando Subespaços Aleatórios e Conjunto de Classificadores

Chu Chia Gean **Celso Antônio Alves Kaestner**

Programa de Pós-Graduação em Informática Aplicada (PPGIA)
Pontifícia Universidade Católica do Paraná (PUCPR)
Rua Imaculada Conceição, 1155 – 80.215-901 – Curitiba – PR – BRASIL

{ccg,kaestner}@ppgia.pucpr.br

Resumo. *Devido à grande quantidade de informação disponível atualmente em meio eletrônico, a tarefa de classificação automática de textos tem ganhado importância nas pesquisas realizadas na área de Recuperação de Informações. Neste artigo é descrita uma nova abordagem para o problema, fundamentada no modelo vetorial para o tratamento de documentos e em técnicas de reconhecimento de padrões.Como as coleções de textos produzem espaços vetoriais de dimensão elevada, o problema foi atacado pelo uso de diversos procedimentos de pré-processamento e por um conjunto de classificadores k-NN (k vizinhos mais próximos), cada um dos quais dedicado a um subespaço do espaço original. A classificação final é obtida pela combinação dos resultados individuais produzidos por cada classificador. Esta abordagem foi aplicada a coleções de documentos extraídas das bases TIPSTER e REUTERS, e os resultados obtidos são apresentados.*

Abstract. *Nowadays, due to the large volume of text available in electronic media, the automatic document classification becomes an important modern Information Retrieval task. In this paper we describe a new approach to the problem, based on the classical vector space model for text treatment and on a Pattern Recognition approach. As texts collections produce huge dimensional vector spaces, we attack the problem using several preprocessing techniques, and a set of k-Nearest-Neighbors classifiers, each of them dedicated to a sub-space of the original space. The final classification is obtained by a combination of the results of the individual classifiers. We apply our approach to a collection of documents extracted from the TIPSTER and REUTERS databases, and the obtained results are presented.*

1. Introdução

Definitivamente vivemos na era da explosão da informação. Estudos recentes divulgados pela Universidade de Berkeley [Lyman 03] indicam que em 2002 foram criados cerca de 5 milhões de *terabytes* de informação em filmes, em meio impresso, ou em meio de armazenamento magnético ou ótico. Este total é equivalente ao dobro do produzido em 1999, o que indica uma taxa de crescimento da ordem de 30 % ao ano. Somente a WWW agrega em torno de 170 *terabytes*, o que equivale a 17 vezes o tamanho das obras impressas da Biblioteca do Congresso dos EUA.

Por outro lado, o uso das informações disponíveis é muito difícil. Diversos problemas tais como a busca de fontes de informação, a recuperação e extração de informações e a classificação automática de textos tornaram-se importantes tópicos de pesquisa em Computação. O uso de ferramentas automáticas para o tratamento de informações tornou-se essencial ao usuário comum; sem eles se torna praticamente impossível desfrutar de todo o potencial informativo disponível na WWW [Zhong 02].

Em particular, a tarefa de classificação automática de documentos reveste-se de importância, visto que é empregada em diversas tarefas cotidianas, tais como a distribuição e seleção automática de *emails* e a classificação de documentos legados [Belkin 92], [Dhillon 01].

Neste artigo propõe-se uma nova abordagem para o problema da classificação automática de textos, com o uso de subespaços vetoriais do espaço original que relaciona termos e documentos, e de classificadores baseados em instâncias (*k*-vizinhos mais próximos) [Mitchell 97] aplicados a estes subespaços. A classificação final é obtida pela combinação dos resultados individuais dos classificadores aplicados aos subespaços.

O enfoque é testado com o auxílio de duas coleções de documentos largamente empregadas para avaliação da tarefa de classificação automática de textos: a base TIPSTER [Trec 04] e a base REUTERS-21578 [Lewis 04].

O restante deste trabalho é organizado da seguinte forma: a seção 2 apresenta uma visão geral do modelo vetorial utilizado para a representação de documentos, e descreve os procedimentos de pré-processamento e a tarefa objetivo. Na seção 3 é apresentado o formalismo subjacente à proposta. A seção 4 descreve a metodologia empregada para a realização dos experimentos e apresenta dos resultados obtidos. Finalmente a seção 5 apresenta algumas conclusões, perspectivas de trabalho e pesquisas futuras.

2. O modelo vetorial e a classificação automática de documentos

No contexto do tratamento de documentos objetivo principal de um modelo de representação é a obtenção de uma descrição adequada da semântica do texto, de uma forma que permita a execução correta da tarefa alvo, de acordo com as necessidades do usuário.

Diversos modelos têm sido propostos, tais como o modelo booleano [Wartik 92], o modelo probabilista [vanRijsberger 92] e o modelo vetorial [Salton 97]. Neste trabalho é utilizado o modelo vetorial, conforme proposto por Salton; no modelo a unidade básica do texto é denominada *termo*, e pode corresponder a uma palavra, a um radical (*stem*) ou a uma sub-cadeia (*substring*) originária do texto, conforme o procedimento de pré-processamento que será detalhado adiante.

De acordo com o modelo vetorial cada documento é modelado por um vetor no espaço *m*-dimensional, onde *m* é o número de diferentes termos presentes na coleção. Os valores das coordenadas do vetor que representa o documento estão associados aos termos, e usualmente são obtidos a partir de uma função relacionada à freqüência dos termos no documento e na coleção.

Pré-processamento

Na etapa de pré-processamento os documentos, considerados aqui como sendo texto "puro", livre de qualquer formato, são tratados de maneira a produzir uma representação mais compacta que seja mais adequada à realização da tarefa objetivo [Sparck Jones 97].

Uma etapa de pré-processamento típica inclui:

1) A eliminação de palavras comuns: as palavras comuns (*stop words*) são elementos de texto que não possuem uma semântica significativa; sua presença não agrega nenhuma indicação do conteúdo ou do assunto do texto correspondente. Normalmente as palavras comuns são constituídas de artigos, preposições, verbos auxiliares, etc, tais como "*the*", "*a/an/one*", "*in*" ou "*is*". Após sua eliminação obtém-se uma representação reduzida do texto, ainda em formato livre.

2) A obtenção dos radicais (*stems*): em linguagem natural diversas palavras que designam variações indicando plural, flexões verbais ou variantes são sintaticamente similares entre si. Por exemplo as palavras "*delete*", "*deletes*", "*deleted*" and "*deleting*" tem sua semântica relacionada. O objetivo da obtenção dos radicais é a obtenção de um elemento único – o radical – que permita considerar como um único termo, portanto com uma semântica única, estes elementos de texto. Este passo permite uma redução significativa no número de elementos que compõem o texto.

Outra possibilidade de pré-tratamento é a obtenção da representação em *n-grams* do texto [Cavnar 94]: constitui-se em uma representação alternativa, onde os termos são obtidos diretamente como sub-cadeias de comprimento *n* das palavras que compõem o texto original. Por exemplo, a partir da palavra "*house*" e considerando $n = 4$, obtém-se as seguintes 4-grams: "*_hou*", "*hous*", "*ouse*" e "*use_*", onde "_" é usado para indicar o início ou fim da palavra.

Evidentemente os procedimentos (1) e (2) acima descritos exigem conhecimentos lingüísticos do idioma em que o documento foi escrito. Já o uso de *n-grams* é completamente independente de idioma.

O pré-processamento pode ainda incluir uma filtragem dos elementos restantes do texto, com base na freqüência com que os mesmos aparecem no documento ou na coleção. O objetivo desta filtragem é o de limitar o número de termos a serem considerados.

Após a etapa de pré-processamento os documentos podem ser considerados como vetores em conformidade com o modelo vetorial. Os termos podem corresponder diretamente aos elementos de texto, aos *stems*, ou às *n-grams*. A dimensão do espaço vetorial total de documentos corresponde ao número de termos considerados em toda a coleção.

Formalmente, seja $C = \{d_1, d_2, \ldots, d_N\}$ uma coleção não-ordenada de documentos d_i, com M diferentes termos. Então a representação de um documentos será $d_i = (f_{i1}, f_{i2}, \ldots, f_{im})$ para $i = 1$ até N, onde f_{ij} é uma função de avaliação associada ao termo j no documento i. A função de avaliação (ou "peso") f_{ij} mais comumente

utilizada no modelo vetorial é conhecida como métrica *tf* * *idf* [Salton 97], na qual: $f_{ij} = tf_{ij} \ln(\frac{N}{idf_j})$, onde tf_{ij} é a freqüência do termo *j* no documento *i* (*term frequency – tf*), idf_j é o número de documentos que contem o termo *j* na coleção (*inter document frequency – idf*), e *N* é o tamanho da coleção (seu número de documentos). Outras medidas, como a freqüência simples (tf_{ij}), também são usadas (ver [Salton 97]).

Portanto, em conformidade com o modelo vetorial uma coleção de documentos pode ser vista como uma imensa matriz $C_{N \times M}$, onde f_{ij} representa o peso do termo *j* no documento *i*, *M* é o número de termos e *N* é o número de documentos na coleção [Berry 99].

$$C = \begin{bmatrix} f_{11}, f_{12}, \ldots, f_{1M} \\ f_{21}, f_{22}, \ldots, f_{2M} \\ \ldots\ldots\ldots\ldots \\ f_{N1}, f_{N2}, \ldots, f_{NM} \end{bmatrix}$$

Classificação de documentos e o classificador *k*-NN

A classificação de documentos pode ser definida sobre o modelo vetorial como um caso especial de um problema de classificação supervisionada no contexto do Reconhecimento de Padrões [Duda 00].

Considera-se que a coleção de documentos tem uma partição implícita. Cada elemento na partição pertence a uma *classe*, formada pelo subconjunto de documentos que compartilham características comuns. Portanto, pode-se considerar a classe como um atributo especial de cada documento. Um *classificador* é um procedimento que determina, a partir de um documento dado, a sua classe.

Um classificador bem conhecido na área do Reconhecimento de Padrões é o *k*-vizinhos mais próximos (*k*-NN) [Duda 00]. Este algoritmo é amplamente utilizado devido à sua simplicidade conceitual e erro conceitualmente limitado. De maneira abreviada um classificador *k*-NN associa a um documento *d* à classe mais freqüente entre as classes dos *k* vizinhos mais próximos de *d* na coleção, de acordo com uma distância calculada no espaço vetorial de documentos.

Na área do tratamento de textos as distâncias entre dois documentos d_i e d_j mais comumente utilizadas são a distância euclidiana $dist(d_i, d_j) = \left[\sum_{k=1}^{M} (f_{ik} - f_{jk})^2\right]^{1/2}$ e a denominada "métrica do co-seno" $cos(d_i, d_j) = \frac{d_i * d_j}{\|d_i\| * \|d_j\|}$ [Salton 97].

3. Subespaços aleatórios e combinação de classificadores

Devido à dimensão elevada do espaço de documentos (*M*), propõe-se neste trabalho a divisão do espaço original em diversos subespaços, cada qual tratado por um classificador específico.

Considere-se o caso de *P* subespaços: inicialmente algumas colunas da matriz de (documentos x termos) *C* são selecionadas aleatoriamente. Se *1,2,…M* são as colunas de

C, seja *X* o subespaço projeção sobre estas colunas; *proj $_X$ (C)* representa a sub-matriz obtida de *C* pela projeção de suas linhas sobre *X*, com dimensão *N* x | *X* |, e *proj $_X$ (d)* é a matriz 1 x | *X* | que corresponde a um documento *d*.

Em cada subespaço gerado desta forma um classificador pode atuar. Nos experimentos constantes deste trabalho foram utilizados subespaços de mesma dimensão (isto é | *X* | é constante para cada subespaço *X*). Em cada *X* empregou-se um classificador *k*-NN fundamentado na métrica do co-seno com o critério usual de classificação do algoritmo. Por exemplo, para *k*=1 segue-se o seguinte critério de classificação: Classe (*d*) = Classe(d_i) onde d_i é tal que $cos(d_i,d) < cos(d_j,q)$ para todo $j <> i$.

Quando se aplica a regra de classificação em cada subespaço, obtem-se *P* possivelmente diferentes classificações. Então se deve decidir a classe de *d* usando um procedimento de decisão que leve em conta os resultados individuais dos diferentes classificadores de 1 até *P*. Usualmente para a combinação de classificadores se emprega o princípio do voto da maioria (*majority vote principle*), isto é, assinala-se ao documento *d* a classe mais freqüente entre as *P* assinaladas individualmente pelos classificadores a *d*.

Além desta regras, neste trabalho empregou-se uma segunda regra de combinação: inicialmente um conjunto com todos os documentos que se constituem nos vizinhos mais próximos a *d* é formado; em seguida determina-se a classe de cada um destes documentos e a mais freqüente é indicada. Este procedimento considera apenas documentos diferentes para calcular a classe final, visto que a formação do conjunto intermediário elimina aparecimentos múltiplos dos documentos, não importando o número de vezes em que os mesmo apareçam nas *P* classificações.

O método delineado acima, com o uso de subespaços vetoriais do espaço original de características e o emprego de combinação de classificadores é uma variante da discriminação estocástica, onde diversos classificadores criados estocasticamente são combinados de forma a aumentar a correção preditiva. Este método tem sido utilizado com sucesso em outros domínios, como por exemplo, no reconhecimento de imagens de dígitos manuscritos [Ho 98].

4. Experimentos realizados e resultados obtidos

Para verificar a aplicabilidade dessa abordagem para a classificação automática de documentos, alguns experimentos preliminares já foram realizados e são descritos a seguir neste trabalho.

Os testes foram realizados utilizando-se duas coleções: (1) a coleção TIPSTER, da conferência TREC [Trec 04], uma competição para a avaliação de sistemas de tratamento automático de documentos; e (2) a coleção REUTERS-21578 [Lewis 04], que foi especificamente construída para a avaliação de sistemas de classificação e é largamente utilizada na literatura da área.

A coleção TIPSTER é formada por milhares de documentos em Inglês (em formato XML), com tamanhos variando de uma a duas linhas até uma ou duas páginas. Os documentos estão agrupados em séries formadas por milhares de elementos. A TREC não possui uma tarefa específica de classificação de documentos; no entanto a

partir da tarefa de recuperação de documentos – quando a partir de uma consulta do usuário deve ser recuperada uma lista ordenada de documentos relevantes – é possível se obter uma partição da coleção em classes: são considerados similares documentos que responder a uma mesma consulta. A indicação da relevância dos documentos em relação às consultas foi feita manualmente por um grupo de especialistas.

Para se obter uma coleção adequada à tarefa de classificação foram selecionados, para experimentos preliminares, 60 documentos que são considerados relevantes para 5 consultas, formando uma coleção equilibrada de 5 classes com 12 elementos cada.

No primeiro experimento os documentos foram pré-processados usando-se a eliminação de palavras comuns e a obtenção dos radicais. A lista de palavras comuns que foram eliminadas foi obtida da BOW Library – CMU e utilizou-se o algoritmo de Porter [Porter 97] para o procedimento de *stemming*. No total foram produzidos 2611 termos, gerando uma matriz $C_{60 \text{ X } 2611}$; os elementos de C foram calculados usando a freqüência simples, isto é, com $f_{ij} = tf_{ij}$.

Dos documentos da base 45 foram utilizados para treinamento e 15 para teste. Foram empregados 30 subespaços aleatórios ($P = 30$), cada um dos quais com dimensão 50 ($| X | = 50$). Em cada subespaço empregou-se um classificador k-NN de funcionamento padrão, usando a métrica do co-seno como medida de similaridade. A combinação dos resultados dos classificadores aplicados aos subespaços foi feita de acordo com as duas regras de combinação já descritas: (1) na primeira delas Classe (d) é a classe mais freqüente retornada pelos classificadores; e (2) Classe(d) é obtida como a classe mais freqüente entre os documentos que constituem os k vizinhos retornados por cada classificador, anteriormente agrupados em um único conjunto.

Os resultados obtidos são sumarizados à Tabela 1, em função dos diferentes valores do parâmetro k. A medida empregada para a avaliação é a *correção*, definida como a porcentagem dos documentos corretamente classificados.

Tabela 1: Correção (em %) segundo os diferentes parâmetros, 1º experimento

k	1ª regra para combinação (*majority vote*)	2ª regra para combinação das classificações
1	50,0	93,3
2	66,7	66,7
3	66,7	60,0

Pode-se observar que, surpreendentemente, os melhores resultados foram obtidos para $k = 1$, e que a segunda regar de combinação de classificadores produz resultados superiores.

No segundo experimento os documentos foram pré-processados utilizando-se a eliminação de palavras comuns e aplicação posterior do processo de obtenção de 4-*grams*. Usou-se a mesma lista de *stop-words* (BOW Library) e um procedimento padrão para obter as 4-*grams* [Cavnar 94]. No total foram produzidos 7027 termos, gerando uma matriz $C_{60 \text{ X } 7027}$, cujos elementos foram obtidos por freqüência simples, como no primeiro experimento. A partição utilizada para treinamento e testes (75 % e 25 %) foi a mesma; também se utilizaram 30 subespaços aleatórios ($P = 30$). Para levar em conta a

maior dimensionalidade do espaço produzido pelas 4-*grams*, empregaram-se subespaços de dimensão 150 ($| X | = 150$). Os classificadores utilizados também foram idênticos aos do primeiro experimento: *k*-NN com uso da métrica de similaridade do co-seno.

Os resultados obtidos são sumarizados à Tabela 2, usando a mesma unidade de avaliação: a taxa de correção na classificação.

Tabela 2: Correção (em %) segundo os diferentes parâmetros, 2º experimento

k	1ª regra para combinação (*majority vote*)	2ª regra para combinação das classificações
1	53,3	66,7
2	53,3	53,3
3	53,3	60,0

Estes resultados são compatíveis com os obtidos no primeiro experimento: aqui novamente a segunda regra de decisão produz resultados superiores.

Em seguida foram realizados experimentos utilizando-se a coleção de documentos REUTERS-21578 [Lewis 04]. Esta base é formada por documentos em XML, permitindo que se indique no corpo do documento as classes ao que o mesmo pertence, segundo diversas classificações. As categorias disponíveis são, por exemplo, <Date>; <Topic>; <Place>; <People>; <Orgs>; <Exchanges>; etc. .

Nos experimentos realizados utilizaram-se somente os 1000 documentos que constituem o primeiro grupo da base em questão, e uma única categoria (<Place>) para a determinação das classes. Neste grupo esta categoria constitui 133 classes, das quais as mais freqüentes são "USA" com freqüência 474, a ausência de informação – que aparece 150 vezes; e a classe "UK", com 50 exemplos. Por outro lado 89 classes possuem um único exemplo neste grupo.

O pré-processamento constitui-se da eliminação de palavras comuns, obtenção de radicais, e exigência do aparecimento do termo em no mínimo dois documentos. Obteve-se assim 3633 termos e conseqüentemente uma matriz $C_{1000 \times 3633}$.

A partição utilizada para treinamento e testes foi de 70 % e 30 %, respectivamente. Foram utilizados 30 subespaços vetoriais (P=30) de dimensão $|X|$ = 1000 cada. Foram efetuados experimentos com a função de ponderação $f_{ij} = tf_{ij}$ (freqüência simples) e também com: $f_{ij} = tf_{ij}\ idf_j$ (métrica *tf*idf*). Os resultados obtidos em termos da taxa de correção são apresentados à Tabela 3.

Tabela 3: Correção (em %) segundo os diferentes parâmetros, 3º experimento

f_{ij}	*k*	1ª regra para combinação (*majority vote*)	2ª regra para combinação das classificações
tf	1	59,7	60,3
tf	2	59,7	59,7
*tf*idf*	1	64,7	63,3
*tf*idf*	2	63,0	60,0

Os resultados obtidos preliminarmente nestes três experimentos são compatíveis com outros experimentos relatados na literatura realizados em condições semelhantes, e podem ser considerados como aceitáveis em diversas aplicações práticas de classificação automática ou semi-automática de documentos.

5. Conclusões e trabalho futuros

Este artigo apresenta uma nova proposta para a realização da tarefa de classificação automática de documentos por meio do uso de subespaços vetoriais do espaço original que relaciona termos e documentos.

Neste trabalho utiliza-se o modelo vetorial para a representação de documentos, de forma que a aplicação da proposta é direta. São empregados conjuntos de classificadores *k* vizinhos mais próximos (*k*-NN) e regras para a combinação dos resultados obtidos individualmente por cada classificador.

Os resultados obtidos, embora preliminares, são encorajadores e indicam a aplicabilidade do método.

Está prevista a realização de novos experimentos para uma melhor avaliação da proposta, nas seguintes direções:

1) Aplicação da proposta a uma coleção de maior envergadura, apara avaliar sua escalabilidade;
2) Avaliação mais detalhada dos efeitos do pré-processamento, incorporando outras combinações relacionadas à eliminação de palavras comuns, obtenção de radicais, obtenção de *n-grams*, e de outros filtros;
3) Realização de testes para avaliar a sensibilidade da arquitetura proposta em relação aos diferentes parâmetros envolvidos, tais como a dimensão do subespaço ($|X|$), e variações no número (P) e no tipo dos classificadores, com uso de árvores de decisão, Naïve-Bayes, e outros algoritmos de classificação [Deb 01], [Mitchell 97]; e
4) Uso de técnicas mais sofisticadas para a seleção dos subespaços a considerar, como o emprego da Análise Semântica Latente (LSA) e suas variações [Deerwester 90], [Zha 98], [Zha 98b].

6. Referências

[Baeza-Yates 99] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley, 1999.

[Belkin 92] Belkin, N.; Croft, W. "Information Filtering and Information Retrieval: Two Sides of the Same Coin". *Communications of the ACM*, Nº 35, pp. 29-38, 1992. .

[Berry 99] Berry, M.; Drmac, Z.; Jessup, E. "Matrices, Vector Spaces, and Information Retrieval", *SIAM Review*, Vol. 41, Nº 2, pp.335-362, 1999.

[Cavnar 94] Cavnar, W. B. "Using An N-Gram-Based Document Representation With a Vector Processing Retrieval Model". In *Proceedings Of TREC-3 (Third Text Retrieval Conference)*. Gaithersburg, Maryland, USA, 1994.

[Deb 01] Deb, K. *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, 2001.

[Deerwester 90] Deerwester, S.; Dumais, S.; Furnas, G.; Landauer, T. "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, Vol. 41, Nº 6, pp. 391-407, 1990.

[Dhillon 01] Dhillon, I.; Modha, D. "Concept Decompositions for Large Sparse Text Data using Clustering". *Machine Learning*, Vol. 42, Nº 1, pp. 143-175, 2001.

[Duda 00] Duda, R.; Hart, P.; Stork, D. *Pattern Classification (2nd. Edition)*, Wiley Interscience, 654 p., 2000.

[Ho 98] Ho, T.K. "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, Nº 8, pp. 832-844, 1998.

[Lewis 04] Lewis, D.D. http://www.daviddlewis.com/resources/testcollections/reuters21578/; acessado em [08/03/2004].

[Lyman 03] Lyman, P. and Varian H.R. (2003). How Much Information. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 acessado em [19/01/2004].

[Mitchell 97] Mitchell, T. *Machine Learning*. McGraw-Hill, 414p., 1997.

[Porter 97] Porter, M.F. "An algorithm for suffix stripping". *Program 14*, 130-137. 1980. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) *Readings in Information Retrieval*. Morgan Kaufmann, pp. 313-316, 1997.

[Salton 97] Salton, G.; Buckley, C. "Term-weighting approaches in automatic text retrieval". *Information Processing and Management 24,* 513-523. 1988. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) *Readings in Information Retrieval*. Morgan Kaufmann, pp. 323-328, 1997.

[Sparck-Jones 97] Sparck-Jones, K.; Willet, P. (Eds.) *Readings in Information Retrieval*. Morgan Kaufmann, 1997.

[Trec 04] http://trec.nist.gov/data.html; acessado em [08/03/2004].

[van Rijsbergen 92] van Rijsbergen, C.J. Probabilistic retrieval revisited. *The Computer Journal*, Vol. 35, No. 3, pp. 291-298, 1992.

[Wartik 92] Wartik, S. "Boolean Operations". *In Information Retrieval: Data Structures and Algorithms*. Frakes, W.B.; Baeza-Yates, R. (Eds.), Prentice Hall, pp. 264-292, 1992.

[Zha 98] Zha, H.; Simon, H. "On Updating Problems in Latent Semantic Indexing". *SIAM Journal of Scientific Computing*, Vol. 21, pp. 782-791, 1999.

[Zha 98b] Zha, H.; Marques, O.; Simon, H. "A Subspace-Based Model for Information Retrieval with Applications in Latent Semantic Indexing". IRREGULAR '98, Berkeley, California, USA, *Lecturer Notes in Computer Science* Nº 1457, Springer Verlag, pp.29-42, 1998.

[Zhong 02] Zhong, N.; Liu, J.; Yao, Y. "In Search of the Wisdom *Web*". *IEEE Computer*, Vol. 35, Nº 1, pp. 27-31, 2002.