Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study

ABDELWADOOD MESLEH
Computer Engineering Department
Balqa' Applied University,
P.O.Box: 15008, Faculty of Engineering technology, Amman 11134.
JORDAN

Abstract: - Feature selection is essential for effective and accurate text classification systems. This paper investigates the effectiveness of six commonly used feature selection methods, Evaluation used an in-house collected Arabic text classification corpus, and classification is based on Support Vector Machine Classifier. The experimental results are presented in terms of precision, recall and Macroaveraged F_1 measure.

Key-Words: - SVM, Feature Selection, Information Gain, CHI, Odd Ratio, GSS, NGL, Mutual Information, Arabic Text Classification, Arabic Text Categorization.

1 Introduction

It is known that the volume of Arabic information available on Internet is increasing. This growth motivates researchers to find some tools that may help people to better managing, filtering and classification these huge Arabic information resources. Text Classification (TC) [1] is the task to classify texts to one of a pre-specified set of categories or classes based on their contents. It is also referred as Text categorization, document categorization, document classification or topic spotting.

TC is among the many important research problems in information retrieval IR, data mining, and natural language processing. It has many applications [2] such as document indexing, document organization, text filtering, word sense disambiguation and web pages hierarchical categorization.

TC has been studied as a binary classification approach (a binary classifier is designed for each category of interest), a lot of TC training algorithms have been reported in binary classification e.g. Naïve Bayesian method [3,4], *k*-nearest neighbors (*k*NN) [4,5,6], support vector machines (SVMs) [7], decision tree [8], etc. On the other hand, it has been studied as a multi classification approach e.g. boosting [9], and multi-class SVM [10,11].

In TC tasks, supervised learning is a very popular approach that is commonly used to train TC systems (algorithms). TC algorithms learn classification patterns from a set of labeled examples, given an enough number of labeled examples (Training Set), and the task is to build a TC model. Then we can use the TC system to predict the category (class) of new (unseen) examples (Testing Set). In many cases, the set of input variables (features) of those examples contains redundant features and do not reveal significant input-output (document-category) characteristics. This is why feature selection techniques are essential to improve classification effectiveness.

The rest of this paper is organized as follows. Section 2 summarizes the Arabic text classification and feature selection related work. Section 3 describes the TC design procedure. Experimental Results are shown in section 4. Section 5 draws some conclusions and outlines future work.

2 Related Work

Most of the TC research is designed and tested for English languages articles. However, some TC approaches were carried out for other European languages such as German, Italian and Spanish [12], and some other were carried out for Chinese and Japanese [13,14]. There is a little TC work [15] that carried out for Arabic articles. To our best knowledge, there is only one commercial automatic Arabic text categorizer referred as "Sakhr Categorizer" [16]. Compared to other languages (English), Arabic language has an extremely rich morphology and a complex orthography; this is one of the main reasons [15,17] behind the lack of research in the field of Arabic TC. However, many machine learning approaches have been proposed to classify Arabic documents: SVMs with CHI square feature extraction method [18,19], Naïve Bayesian method [20], k-nearest neighbors (kNN) [21,22,23], maximum entropy [17,24], distance based classifier [25,26,27], Rocchio

algorithm [23] and WordNet knowledge based [28]. It is quit hard to fairly compare the effectiveness of these approaches because of the following reasons:

1. Their authors have used different corpora (because there is no publicly available Arabic TC corpus).
2. Even those who have used the same corpus, it is not obvious whether they have used the same documents for training/testing their classifiers or not.
3. Authors have used different evaluation measures: accuracy, recall precision and F_1 measures.

For English language TC tasks, the valuable studies [29,30] have presented an extensive empirical study of many FS methods with *k*NN and SVMs, it has been reported that CHI and IG [29] performed most effective with kNN classifier. On the other hand, it has been shown that MI and TS [29] performed terribly. However, IG [30] is the best choice to improve SVMs classifier performance in term of precision.

To our best knowledge, the only work that investigated the usage of some FS methods for Arabic language TC tasks is [23], FS methods (IG, CHI, DF, OR, GSS and NGL) have been evaluated using a hybrid approach of light and trigram stemming. In [23], it has been shown that the usage of any of those methods separately gave near results, NGL performed better than DF, CHI and GSS with Rocchio classifier in term of F_1 measure (it was noticed that when using IG and OR, the majority of documents contain non of the selected terms). [23] has concluded that a hybrid approach of DF and IG is the a preferable FS method with Rocchio classifier. It is clear that authors of [23] have not reported the comparison results of the mentioned FS methods in term of recall, precision and F_1 measure, and they have not considered SVMs which was already known to be superior to the classifiers they have studied. In this paper, we have restricted our study of TC on binary classification methods and in particular to SVMs and only for Arabic language articles. On the other hand, through fair comparison experiments, we have investigated the performance of the well known FS methods with SVMs for Arabic language TC tasks.

3 TC Process

TC system design usually compromises the following three main phases [7]: *Data pre-processing and feature selection* phase is to make the text documents compact and applicable to train the text classifier, *text classifier* phase, the core TC learning algorithm, shall be constructed, learned and tuned using the compact form of the Arabic dataset, and *evaluation phase* (using some performance measures). Then the TC system can implement the function of document classification.

The following subsections are devoted to Arabic dataset preprocessing, feature selection methods, text classifier and TC evaluation measures.

3.1 Arabic Dataset Preprocessing

Since there is no publicly available Arabic TC corpus to test our classifier, we have used an in-house collected corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and Al-Dostor as well as a few other specialized websites. The collected corpus contains 1445 documents that vary in length. These documents fall into Nine classification categories that vary in the number of documents (Computer, Economics, Education, Engineering, Law, Medicine, Politics, Religion and Sports). In this Arabic dataset, each document was saved in a separate file within the corresponding category's directory, i.e. this dataset documents are single-labeled.

Arabic documents are processed according to the following steps [5,11,28]:

1. Each article in the Arabic dataset is processed to remove digits and punctuation marks.
2. We have followed [15] in the normalization of some Arabic letters: we have normalize letters "ء" (hamza), " آ" (aleph mad), "أ " (aleph with hamza on top), "ؤ " (hamza on w), "إ " (alef with hamza on the bottom), and "ئ" (hamza on ya) to "ا" (alef). The reason for this normalization is that all forms of hamza are represented in dictionaries as one form and people often misspell different forms of aleph. We have normalized the letter "ى" to "ي" and the letter "ة" to "ه". The reason behind this normalization is that there is not a single convention for spelling "ى" or "ي"and "ة" or "ه" when they appears at the end of a word.
3. All the non Arabic texts were filtered.
4. Arabic function words (such as "آخر", "أبدا", "أحد" etc.) were removed. The Arabic function words (stop words) are the words that are not useful in IR systems e.g. pronouns and prepositions.
5. The vector space representation [31] is used to represent the Arabic documents. In vector space model (VSM), term frequency *TF* concerns with the number of occurrences a term i occurs in document j while inverse document frequency

IDF concerns with the term occurrence in a collection of texts and it is calculated by $IDF(i) = \log(N / DF(i))$, Where N is the total number of training documents and DF is the number of documents that term i occurs in. Using VSM in [32] shown that combining *TF* and *IDF* to weight terms ($IDF.TF$) gives better performance. In our Arabic dataset, each document feature vector is normalized to unit length and the $IDF\,TF$ is calculated.

6. We have not done stemming, because it is not always beneficial [33] for TC tasks. And because it has been empirically proved [18,19] that it is not beneficial for Arabic TC tasks too (this is because the same Arabic root, depending on the context, may be driven from more than one Arabic words.

3.2 Feature Selection Methods

Feature selection (FS) is a process that chooses a subset from the original feature set according to some criterions, it is been widely applied to TC tasks [2,34,35,36,37,38].

FS basic steps are [39]:

1. Feature generation: in this step, a candidate subset of features is generated by some search process.
2. Feature evaluation: using some evaluation criterion, the candidate feature subset is evaluated. (This step measures the goodness of the produced features).
3. Stopping: using some stopping criterion, decide whether to stop or not, i.e. whether a predefined number of features are selected or whether a predefined number of iterations is reached.
4. Feature Validation: using a validation procedure, a decision is made whether a feature subset is valid or not. (As a matter of fact, this step is not a part of FS process itself, but in practice, we need to verify the validity of the FS outcome).

Generally, FS algorithms are commonly accomplished [40] by a filter-based method which selects a subset of features by filtering based on the scores which were assigned by a specific weighting method, by a wrapper approach, where the subset of features is chosen based on the accuracy of a given classifier or by a hybrid method which takes advantage of the filter and wrapper methods. The major disadvantage of wrapper methods is its computational cost, this makes wrapper methods impractical for large classification problem. Instead filter methods are often used.

In TC task, because the number of features is huge, an important consideration shall be made to select the right FS method to improve the performance of the TC task in terms of learning speed and effectiveness, to reduce data dimension and remove irrelevant, redundant, or noisy data. On the other hand, FS may decrease accuracy (over-fitting problem [1], which may arise when the number of features is large and the number of training samples is relatively small).

In addition to classical FS methods [29] (Document frequency thresholding (DF), The X^2 statistics (CHI), Term strength (TS), Information gain (IG) and Mutual information (MI)), Other FS methods have been reported in literatures such as Odds Ratio [41], NGL [42], GSS [43], etc.

Table 2 contains the functions for commonly used FS methods [2], where t_k denotes a term, c_i denotes a category. DF for a term t_k is the number of documents in which t_k occurs, probabilities are interpreted on events of training document space, for example $P(t_k, \bar{c}_i)$ denotes the probability that a term t_k occurs in a document x that does not belong to class c_i, $P(\bar{c}_i)$ is estimated as the number of documents that do not belong to class c_i divided by the total number of training documents.

In this paper, we have restricted our study on only six FS methods and in particular to CHI, NGL, GSS, IG, OR and MI FS methods.

Table 2: Commonly used FS Methods.

CHI	$\frac{N.[P(t_k,c_i).P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$
NGL	$\frac{\sqrt{N}.[P(t_k,c_i).P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)]}{\sqrt{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}}$
GSS	$P(t_k,c_i).P(\bar{t}_k,\bar{c}_i) - P(t_k,\bar{c}_i)P(\bar{t}_k,c_i)$
IG	$\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_i,\bar{t}_i\}} P(t \mid c).\log\frac{P(t \mid c)}{P(t).P(c)}$
OR	$\frac{P(t_k \mid c_i).(1-P(t_k \mid \bar{c}_i))}{(1-P(t_k \mid c_i)).P(t_k \mid \bar{c}_i)}$
MI	$\log\frac{P(t_k,c_i)}{P(t_k).P(c_i)}$

3.2 Text Classifier:

SVMs based classifiers are binary classifiers, which are originally proposed by [44]. Based on the structural risk minimization principle, SVM seeks a decision hyperplane to separate the training data points into two classes and makes decisions based on the support vectors that are carefully selected as

the only effective elements in the training data set. In the non-separable case, the optimization of SVM is to minimize equation (1).

$$\begin{aligned} &\min.\ \frac{1}{2}|w|^2 + C\sum_i \xi_i, \\ &s.t.\ \forall i, y_i(x_i w + b) - 1 + \xi_i \geq 0, \quad (1) \\ &\qquad \forall i, \xi_i \geq 0. \end{aligned}$$

3.3 TC Evaluation Measures:

Text classification performance is always evaluated in terms of categorization effectiveness [45] which is measured in terms of precision, recall and F_1 measure. Denote the precision, recall and F_1 measure for a class C_i by P_i, R_i and F_i, respectively:

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad R_i = \frac{TP_i}{TP_i + FN_i}, F_i = \frac{2P_i R_i}{R_i + P_i}$$

Where: TP_i: true positives; the set of documents that both the classifier and the previous judgments (as recorded in the test set) classify under C_i, FP_i: false positives; the set of documents that the classifier classifies under C_i, but the test set indicates that they do not belong to C_i. TN_i: true negatives; both the classifier and the test set agree that the documents in TN_i do not belong to C_i. FN_i: false negatives; the classifier does not classify the documents in FN_i under C_i, but the test set indicates that they should be classified under C_i.

4 TC Experimental Results

In our experiments, we have used the mentioned Arabic dataset for training and testing our Arabic text classifier. In addition to the mentioned preprocessing steps in section 3, we have filtered all terms with term frequency *TF* less than some threshold (threshold is set to Three for positive features and set to Six for negative features in training documents). We have used an SVM package, TinySVM (downloaded from http://chasen.org/~taku/), the soft-margin parameter C is set to 1.0 (other values of C shown no significant changes in results). First of all, we have conducted a classification experiment without feature selection where all the 78699 terms were selected. Then to fairly compare the six FS methods (CHI, NGL, GSS, IG, OR and MI), we have conducted three groups of experiments. For each group and for each text category, we have randomly specified one third of the articles and used them for testing while the remaining articles used for training the SVM classifier. And for each FS method, we have conducted three experiments: the first experiment selects the 180 top features, the second experiment selects the 160 top features and finally the third experiment selects the 140 top features. The results are shown in Figure 1. We conclude that CHI, NGL and GSS performed most effective with SVMs for Arabic TC tasks, but OR and MI performed terribly.

5 Conclusion

We have investigated the performance of six FS methods with SVMs evaluated on an Arabic dataset. CHI square performance is best. In future, we like to study more FS methods for our SVMs based Arabic TC system. And we like to deeply investigate the effect of the FS methods on small categories (such as C*omputer*).

References:

[1] C. Manning, and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press (1999).

[2] F. Sebastiani, Machine Learning in Automated Text Categorization, *ACM Computing Surveys*, Vol. 34, No. 1, 2002, pp.1-47.

[3] A. McCallum, and K. Nigam, A comparison of event models for naïve Bayes text classification, *AAAI-98 Workshop on Learning for Text Categorization*, 1998, pp.41-48.

[4]Y. Yang, and X. Liu, A re-examination of text categorization methods, *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*,1999, pp. 42-49.

[5]G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, An kNN Model-based Approach and its Application in Text Categorization, *Proceeding of 5th International Conference on Intelligent Text Processing and Computational Linguistic, CICLing-2004, LNCS 2945*, Springer-Verlag, pages, 2004, pp. 559-570.

[6]Y.M. Yang, "Expert network: Effective and efficient learning from human decisions in text categorization and retrieval", Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 13-22.

[7] T. Joachims, Text categorization with Support Vector Machines: learning with many relevant features, *Proceedings of the European*

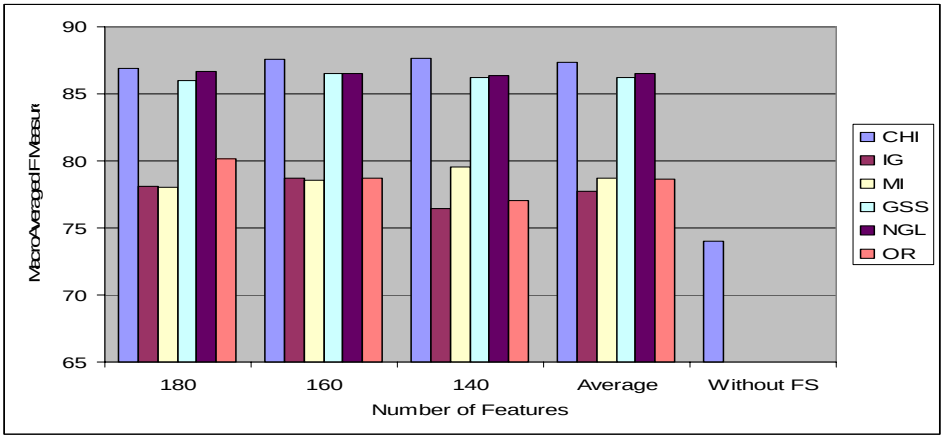

Figure 1: MacroAveraged F_1 Measure for SVM with Six FS Methods at Different Sizes of Features.

Conference on Machine Learning, Berlin, 1998, pp.137-142, Springer.

[8]D. Lewis, and M. Ringuette, A comparison of two learning algorithms for text categorization, *the 3rd annual Symposium on Document Analysis and Information Retrieval*, 1994, pp.81-93.

[9]R. Schapire, and Y. Singer, BoosTexter: A boosting-based system for text categorization, *Machine Learning*, Vol. 39, No.2-3, 2000, pp.135-168.

[10]S. Gao, W. Wu, C-H. Lee, and T-S. Chua, A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization, *ACM Transactions on Information Systems*, Vol. 24, No. 2, 2006, pp. 190-218.

[11]J. Zhang, R. Jin, Y.M. Yang, and A. Hauptmann, A modified logistic regression: an approximation to SVM and its applications in large-scale Text Categorization, *Proceedings of the Twentieth International Conference*, August 21-24, 2003, Washington, DC, USA, pp. 888-895.

[12]F. Ciravegna, et.al., Flexible Text Classification for Financial Applications: the FACILE System, *Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000*, 2000, pp.696-700.

[13]F. Peng, X. Huang, D. Schuurmans, and S. Wang, Text Classification in Asian Languages without Word Segmentation, *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages, Association for Computational Linguistics*, July 7, Sapporo, Japan, 2003, pp. 41-48.

[14]J. He, A-H. TAN, and C-L. TAN, On Machine Learning Methods for Chinese document Categorization, *Applied Intelligence*, 2003, pp. 311-322.

[15]A.M. Samir, W. Ata, and N. Darwish, A New Technique for Automatic Text Categorization for Arabic Documents, *Proceedings of the 5th Conference of the Internet and Information Technology in Modern Organizations*, December, Cairo, Egypt, 2005, pp. 13-15.

[16] Sakhr Company: http://www.sakhr.com.

[17]A.M. El-Halees, Arabic Text Classification Using Maximum Entropy, *The Islamic University Journal*, Vol. 15, No. 1, 2007, pp 157-167.

[18]A.M. Mesleh, CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System, *Proceedings of the 2nd International Conference on Software and Data Technologies, (Knowledge Engineering),* Vol. 1, Barcelona, Spain, July, 22—25, 2007, pp. 235-240.

[19]A.M. Mesleh, CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System, *Journal of Computer Science*, Vol. 3, No. 6, 2007, pp. 430-435.

[20]M. Elkourdi, A. Bensaid, and T. Rachidi, Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm, *Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, August 23rd-27th , 2004, pp. 51-58.

[21]G. Kanaan, R. Al-Shalabi, and A. AL-Akhras, KNN Arabic Text Categorization Using IG Feature Selection, *Proceedings of The 4th International Multiconference on Computer Science and Information Technology*, Vol. 4, Amman, Jordan, April 5-7, 2006.

[22]R. Al-Shalabi, G. Kanaan, M. Gharaibeh, Arabic Text Categorization Using kNN Algorithm, *Proceedings of The 4th International Multiconference on Computer Science and Information Technology,* Vol. 4, Amman, Jordan, April 5-7, 2006.

[23]M. Syiam, Z. Fayed, and M. Habib, An Intelligent System for Arabic Text Categorization, *International Journal of Intelligent Computing and Information Sciences*, Vol.6, No.1, 2006, pp. 1-19.

[24]H. Sawaf, J. Zaplo, and H. Ney, Statistical Classification Methods for Arabic News Articles, *Paper presented at the Arabic Natural Language Processing Workshop (ACL2001),* Toulouse, France. (Retrieved from Arabic NLP Workshop at ACL/EACL 2001 website: http://www.elsnet.org/acl2001-arabic.html).

[25]R.M. Duwairi, A Distance-based Classifier for Arabic Text Categorization, *Proceedings of the 2005 International Conference on Data Mining,* Las Vegas, USA, 2005, pp.187-192.

[26] L. Khreisat, Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study, *Proceedings of the 2006 International Conference on Data Mining.* Las Vegas, USA, 2006, pp.78-82.

[27]R.M. Duwairi, Machine Learning for Arabic Text Categorization, *Journal of American society for Information Science and Technology*, Vol. 57, No. 8, 2006, pp.1005-1010.

[28]M. Benkhalifa, A. Mouradi, and H. Bouyakhf, Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization, *International Journal of Intelligent Systems*, Vol. 16, No. 8, 2001, pp. 929-947.

[29]Y.M. Yang, and J.O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, *In J. D. H. Fisher, editor, The 14th International Conference on Machine Learning*, Morgan Kaufmann, 1997, pp.412-420.

[30]G. Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1289-1305.

[31]G. Salton, A. Wong, and C.S. Yang, A Vector Space Model for Automatic Indexing, *Communications of the ACM*, Vol. 18, No. 11, 1975, pp. 613-620.

[32]G. Salton, and C. Buckley, Term weighting approaches in automatic text retrieval, *Information Processing and Management*, Vol. 24, No. 5, 1988, pp. 513-523.

[33]T. Hofmann, Introduction to Machine Learning, Draft Version 1.1.5, November 10, 2003.

[34]E. Leopold, and J. Kindermann, Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?, *Machine Learning*, Vol. 46, 2002, pp. 423-444.

[35]K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell, Text Classification from Labeled and Unlabeled Documents Using EM, *Machine Learning*, Vol. 39, 2000, pp. 103-134.

[36]D. Mladenic, Feature subset selection in text learning, *Proceedings of European Conference on Machine Learning,* 1998, pp. 95-100.

[37]H. Taira, and M. Haruno, Feature selection in SVM text categorization, *Proceedings of AAAI-99, 16th Conference of the American Association for Artificial Intelligence* (Orlando, US, 1999), 1999, pp. 480-486.

[38]D.Lewis, Feature Selection and Feature Extraction for Text Categorization, *Proceedings of a workshop on speech and natural language*, San Mateo, CA: Morgan Kaufmann, 1992, pp. 212-217.

[39]H. Liu, and L. Yu, Toward integrating feature selection algorithms for classification and clustering, *IEEE Transaction on Knowledge and Data Engineering*, Vol. 17, No. 4, 2005, pp. 491-502.

[40]H. Liu, Evolving feature selection, *IEEE Intelligent Systems*, 2005, pp.64-76.

[41]D. Mladenic, and M. Grobelnik, Feature Selection for Unbalanced Class Distribution and Naïve Bayes, *Proceedings of the 16th International Conference on Machine Learning*, 1999, pp. 258-267.

[42]H.T. Ng, W.B. Goh, and K.L. Low, Feature Selection, Perceptron Learning, and A usability Case Study for Text Categorization, *Proceedings of SIGIR-97, 20th ACM International Conference on Research and Development in Information Retrieval*, Philadelphia, PA, 1997, pp. 67-73.

[43]L. Galavotti, F. Sebastiani, and M. Simi, Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization, *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*, Lisbon, Portugal, 2000, pp. 59-68.

[44]V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[45]R. Baeza-Yates, and B. Rieiro-Neto, *Modern Information Retrieval*, Addison-Wesley & ACM Press, 1999.