ORIGINAL RESEARCH

A text feature selection method based on category-distribution divergence

Yonghe Lu*, Wenqiu Liu, Xinyu He

School of Information Management, Sun Yat-sen University, Guangzhou, China

Received: July 11, 2015 **Accepted:** August 13, 2015 **Online Published:** September 7, 2015
DOI: 10.5430/air.v4n2p143 **URL:** http://dx.doi.org/10.5430/air.v4n2p143

Abstract

The purpose of this paper is to overcome the problem that traditional feature selection methods [such as document frequency(DF), Chi-square statistic(CHI), information gain(IG), mutual information(MI) and Odds ratio(OR)] do not consider the distribution of features among different categories. The work aims at selecting the features that can accurately represent the theme of texts and to improve the accuracy of classification. In this paper, we propose a text feature selection method based on Category-Distribution Divergence, and the degree of membership and degree of non-membership are introduced into CDDFS (feature selection based on category-distribution divergence). CDDFS is used as a filter which can filter the features having low degree of membership and high degree of non-membership. CDDFS is tested with five feature selection methods and three classifiers using the corpus of Sogou Lab Data, and experimental results show that this method performs better than other feature selection methods when using KNN, and close to CHI when using Rocchio algorithms and SVM at high dimensions. This research proposes the representativeness and distinguishability of feature for category, and the representativeness and distinguishability of feature for non-category. If a feature has good distinguishability and high representativeness, then this feature will be retained in feature selection.

Key Words: Text categorization, Category-distribution, Feature selection, Vector space model

1 Introduction

Text categorization is a key technique that aims at processing and organizing large amounts of text data, and it can solve some problems brought by the rapid growth of information. In text categorization, data dimension has a direct impact on the results and speed of categorization. For most categorization algorithms, the high dimensional data (especially the data with thousands even ten thousands dimensions) will make the classifier stop working due to excessive computing or consuming too many resources.[1–3]

Traditional feature selection methods are document frequency (DF), Chi-square statistic (CHI), information gain (IG), mutual information (MI) and Odds ratio (OR) and so on. Although these methods can reduce the dimensions of feature vector, simplify the calculation and reduce the training time of models, they are also existing some disadvantages.[4–7] To overcome their shortcomings, many scholars proposed different modified methods. Liu *et al.*[6] put forward a kind of optimizing MI text feature selection method. This method can improved the efficiency of MI model. Considering that traditional Information Gain ignores the shortcoming of distributing information inside class and between classes, Guo and Liu[8] proposed modified IG by introduce the distribution information inside class and concentration information between classes. And the experimen-

*Correspondence: Yonghe Lu; Email: luyonghe@mail.sysu.edu.cn; Address: School of Information Management, Sun Yat-sen University, Guangzhou, China.

tal results verify the efficiency and probability of the improved IG approach. Liu *et al.*[9] proposed a novel hybrid method MRMR (maximum relevance minimum redundancy) to select features and to improve the accuracy of the ELM (extreme learning machine) classifier. Pinheiro *et al.*[10] proposed two filtering methods for feature selection, namely: MFD (Maximum Features per Document) and MFDR (Maximum Features per Document-Reduced). Lee and Kim[11] proposed the Mutual Information-based multi-label feature selection method using interaction information. Furthermore, scholars also proposed some modified algorithms based on other areas of knowledge.[12–15] For example, Ghamisi and Benediktsson[16] proposed a feature selection approach based on the integration of a genetic algorithm and particle swarm optimization. LIN *et al.*[17] introduced a novel method employing domain ontology to extract feature. Xu *et al.*[18] put forward a new feature selection function KG (knowledge gain).

For the existing approach do not consider features' distribution divergence among different categories, we propose a new feature selection method–feature selection based on category-distribution divergence (CDDFS). And the degree of membership and degree of non-membership are introduced into CDDFS. Firstly, we calculate the degree of membership between a word and its category. Secondly, we calculate the degree of non-membership between a word and other categories. Lastly, we combine the degree of membership and degree of non-membership together.

2 Feature selection based on category-distribution divergence

2.1 The related theory

We assume there are three features t_i, t_j and t_k. If the feature t_i usually appears in the category C, and t_j appears in all categories with almost equal frequency, then t_i has higher category distinguishability than t_j. We assume that both t_i and t_k have high category distinguishability in category C. If the occurrence frequency of t_i is higher than t_k, then we believe t_i has higher category representation than t_k in category C. For example, "Money" and "The SFC" (also known as The Securities and Futures Commission) appear frequently in the category "Finance". Obviously, they both have high category distinguishability of "Finance", but the occurrence frequency of "Money" is higher than "The SFC". At this time, we believe that "Money" has higher category representation than "The SFC" in category "Finance".

Furthermore, if a feature t has a low frequency in category C but a high frequency in other categories (non-category C), it means that t has a high distinguishability for non-category C. And the appearance of t can help to judge that the text do not belong to category C. Besides, in non-category C, the higher frequency of feature t has, the stronger representativeness of feature t has. For instance, "currency" and "stock exchange" have a very low frequency in category "Sport" while they have a high frequency in non-category "Sport". It means that these two features have a high distinguishability to non-category "Sport". Moreover, if "currency" has a higher frequency than "stock exchange" in non-category "Sport", then "currency" is considered to have a stronger representativeness than "stock exchange" to non-category "Sport".

To sum up, for any categories, the representativeness and distinguishability of feature to category are called degree of membership. Meanwhile, the representativeness and distinguishability of feature to non-category are called degree of non-membership. For any feature in a category, if the feature has high degree of membership and low degree of non-membership, which is meant that this feature has good distinguishability and high representativeness, then this feature should be retained in feature selection.

2.2 Construct the function

We define that t_i is the i-th feature, (t_i) - is "NOT" operation to t_i. c_j is the j-th category, $(\bar{c_j})$ is "NOT" operation to c_j. N_{11} refers to the number of texts which contain feature t_i and belong to category c_j. N_{10} refers to the number of texts which contain t_i and do not belong to c_j. N_{01} refers to the number of texts which do not contain t_i but belong to c_j. N_{00} refers to the number of texts which neither contain t_i nor belong to c_j (see Table 1).

Table 1: The Relationship among Parameters

	c_j	$\bar{c_j}$
t_i	N_{11}	N_{10}
$\bar{t_i}$	N_{01}	N_{00}

Based on the above reasons, we define M as the total number of categories and N as the total number of texts. The distinguishability of feature t_i to category c_j is computed as Equation (1).

$$diff(t_i, c_j) = \log_2(\frac{N_{11}}{N_{10}+1} + 1) \tag{1}$$

The representativeness of t_i to c_j is computed as Equation (2).

$$repr(t_i, c_j) = N_{11} \tag{2}$$

Then the degree of membership of t_i to c_j is given by Equation (3).

$$\begin{aligned} belong_{positive}(t_i, c_j) &= diff(t_i, c_j) * repr(t_i, c_j) \\ &= N_{11}\log_2(\frac{N_{11}}{N_{10}+1} + 1) \end{aligned} \tag{3}$$

Similarly, the degree of non-membership of t_i to c_j (the degree of membership of t_i to $(\bar{c_j})$) is defined by Equation (4).

$$belong_{positive}(t_i, c_j) = N_{10}\log_2(\frac{N_{11}}{N_{10}+1}+1) \quad (4)$$

Considering the degree of membership and non-membership of t_i to c_j, we can get the CDDFS with equation (5).

$$\begin{aligned} belong(t_i, c_j) =& N_{11}\log_2(\frac{N_{11}}{N_{10}+1}+1) \\ &- N_{10}\log_2(\frac{N_{10}}{N_{11}+1}+1) \end{aligned} \quad (5)$$

If belong$(t_i, c_j) \leq 0$, that means the degree of membership of t_i to c_j is lower than the degree of non-membership. Then the feature t_i should not be selected in feature selection. The evaluation function of CDDFS is computed by equation (6).

$$\begin{aligned} &evaluation_{CDDFS}(t_i, c_j) \\ &= \begin{cases} \log_2(belong(t_i, c_j)+1) & belong(t_i, c_j) > 0 \\ 0 & belong(t_i, c_j) \leq 0 \end{cases} \end{aligned} \quad (6)$$

3 Experimental results and analysis

3.1 Experimental setting

To verify the validity of feature selection algorithm CDDFS, we compare it with DF,[19] CHI, IG, MI and OR. The data set is Sogou corpus of text categorization.[20] We select nine categories from the corpus, namely automotive, finance, IT, health, sports, tourism, education, recruitment and military. We select 200 texts for each category, and these texts are divided into training set and testing set according the ratio of 1:1. There are 900 texts in both training set and testing set respectively. We use open source package Lucene[21] to preprocess the text set, including Chinese word segmentation, word frequency statistic and so on. The document representation model is Vector Space Model (VSM), and we do experiments at 360, 720, $\cdots$, 3600 dimensions respectively.

The value of Global evaluation function of feature t_i is computed by equation (7).

$$evaluation(t_i) = \max_{j=1\cdots M}\{evaluation(t_i, c_j)\} \quad (7)$$

Where M is the number of categories. The feature weight calculation method is traditional TF-IDF, and it is computed by equation (8).

$$w_{id} = tf_{id}\log\frac{D}{df_i} \quad (8)$$

Where tf_{id} is the number of feature i appearing in text d. D is the text number of training set. df_i is the number of texts which contain feature i in training set.

In order to eliminate the influence of document's length on categorization results, we use the cosine normalization.[22] It is calculated in equation (9).

$$w_{id} = \frac{tf_{id}\log\frac{D}{df_i}}{\sqrt{\sum_{i=1}^{N}(tf_{id}\log\frac{D}{df_i})^2}} \quad (9)$$

We select KNN, center-point method and SVM as the classifier. The k in KNN is 7, and the similarity is computed by equation (10).

$$sim(d_i, d_j) = \cos\alpha = \frac{\sum_{k=1}^{n}(w_{ik}w_{jk})}{\sqrt{\sum_{k=1}^{n}w_{ik}^2\sum_{k=1}^{n}w_{jk}^2}} \quad (10)$$

Where $sim(d_i, d_j)$ is the similarity between text d_i and d_j, w_{ik} is the k-th feature weight in text d_i, n is the dimension number of feature vector.

With the Cross-validation method, we get the parameters in SVM. They are cost=8, gamma=0.38125. The evaluation criteria of categorization results are MR (macro-averaging recall), MP (macro-averaging precision) and MF (macro-averaging f-measure).

In order to verify the classification performance with CDDFS is significantly better than the ones using other feature selection methods, paired-sample T-test was used for significance test.

Before using paired-sample T-test, we set up two hypotheses. The first was the null hypothesis, which assumed that the mean of two paired samples were equal. The second hypothesis was an alternative hypothesis, which assumed that the means of two paired samples were significantly different. And we choose a significance level of 0.05, it meant there was a 5% chance of rejecting the null hypothesis when it was true.

3.2 Experimental results

After using various feature selection methods and three different classifiers, we get the macro-averaging F1 of the categorization results, and they are shown in Figures 1-3 and Tables 2-4.

3.2.1 The experiment with KNN classifier

From Figure 1 and Table 2, we can know that CDDFS has better classification performance than other feature selections when using KNN as classifier.

This experiment used six different feature selection methods. Table 2 describes MacF$_1$ in different features dimensions for six feature selections. Table 3 shows the comparison of MacF$_1$ by using paired T test.

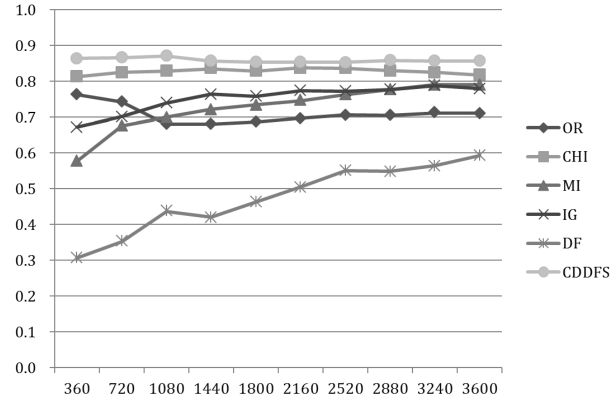

Figure 1: Experimental Results with KNN

Table 2: Experimental Results with KNN

Dimension Number	OR	CHI	MI	IG	DF	CDDFS
360	0.762397	0.813257	0.577079	0.671722	0.307134	0.863770
720	0.742256	0.825844	0.675333	0.701573	0.352370	0.865793
1080	0.680397	0.828759	0.699871	0.740211	0.437025	0.869975
1440	0.680405	0.835206	0.722009	0.763695	0.420384	0.857111
1800	0.686937	0.828902	0.734031	0.757990	0.463555	0.853632
2160	0.697051	0.837103	0.746118	0.774204	0.505097	0.854226
2520	0.706218	0.836109	0.763378	0.772585	0.550143	0.853348
2880	0.705218	0.830364	0.777507	0.776194	0.548237	0.858916
3240	0.711690	0.825478	0.790843	0.787653	0.564861	0.856722
3600	0.711022	0.817522	0.790695	0.779897	0.592975	0.856823

Table 3 shows that all comparisons of $Mac\mathrm{F}_1$ seem to reject the null hypothesis, thus demonstrating the CDDFS is better than the other feature selections when using KNN as classifier.

Table 3: Comparison of $Mac\mathrm{F}_1$ by using paired t test

	t	*p*
OR *vs.* CDDFS	-18.876	.000
CHI *vs.* CDDFS	-8.745	.000
MI *vs.* CDDFS	-6.040	.000
IG *vs.* CDDFS	-8.064	.000
DF *vs.* CDDFS	-12.321	.000

3.2.2 *The experiment with Rocchio Algorithms*

As we can see from the Figure 2 and Table 4, when the dimensions are less than 2,520, CDDFS have better performance than other feature selections when using Rocchio algorithms. When the dimensions are more than 2,520, CDDFS is as good as CHI and its performance is better than OR, MI, IG, DF.

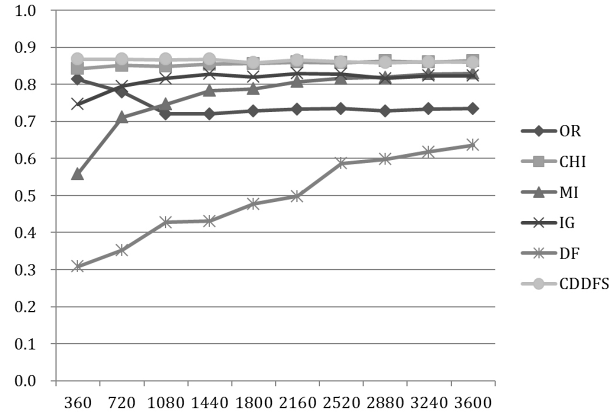

Figure 2: Experimental Results with Rocchio Algorithms

Table 4: Experimental Results with Rocchio Algorithms

Dimension Number	OR	CHI	MI	IG	DF	CDDFS
360	0.812725	0.842178	0.558861	0.746689	0.307942	0.866788
720	0.778864	0.851801	0.711011	0.795018	0.353030	0.867351
1080	0.720103	0.848168	0.746284	0.815820	0.428336	0.866494
1440	0.719991	0.854049	0.782377	0.827047	0.430495	0.867227
1800	0.728231	0.855923	0.787387	0.819956	0.477501	0.857714
2160	0.732612	0.859613	0.807248	0.829748	0.497927	0.865251
2520	0.735075	0.857378	0.815842	0.827985	0.586166	0.861749
2880	0.728048	0.862398	0.819616	0.816466	0.597405	0.857228
3240	0.733851	0.860097	0.828310	0.822511	0.617944	0.860439
3600	0.734742	0.863981	0.829789	0.822929	0.636366	0.859246

This experiment used six different feature selection methods. Table 4 is the value of $Mac\mathrm{F}_1$ in different features dimensions for six feature selections. Table 5 shows the comparison of $Mac\mathrm{F}_1$ by using paired T test.

Table 5 shows that all comparisons of $Mac\mathrm{F}_1$ seem to reject the null hypothesis, thus demonstrating the CDDFS is better than the other feature selections when using Rocchio algorithms as classifier.

Table 5: Comparison of $Mac\mathrm{F}_1$ by using paired t test

	t	*p*
OR *vs.* CDDFS	-13.458	.000
CHI *vs.* CDDFS	-2.319	.046
MI *vs.* CDDFS	-3.488	.007
IG *vs.* CDDFS	-5.924	.000
DF *vs.* CDDFS	-9.932	.000

3.2.3 *The experiment with SVM classifier*

As we can see from the Figure 3 and Table 6, CDDFS offers a better performance than other feature selections except CHI when using SVM. Although CHI has the best performance, CDDFS is easier to understand and its calculation is simpler than CHI.

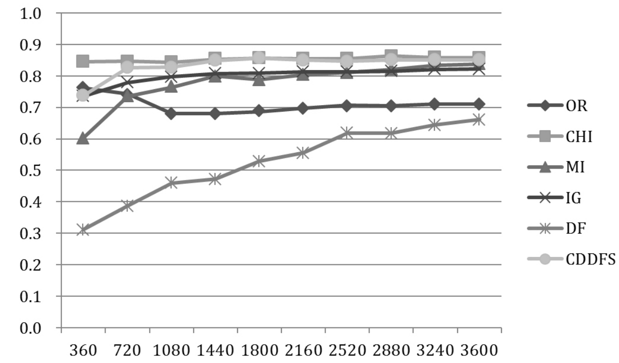

Figure 3: Experimental Results with SVM

Table 6: Experimental Results with SVM

Dimension Number	OR	CHI	MI	IG	DF	CDDFS
360	0.762397	0.845131	0.602185	0.736102	0.312761	0.736803
720	0.742256	0.846951	0.735232	0.779516	0.387001	0.826019
1080	0.680397	0.842320	0.764729	0.797358	0.460074	0.827735
1440	0.680405	0.853607	0.800115	0.807801	0.472918	0.848906
1800	0.686937	0.855924	0.788876	0.808315	0.528923	0.856675
2160	0.697051	0.855088	0.803935	0.813172	0.555581	0.850049
2520	0.706218	0.854340	0.810758	0.813137	0.618924	0.846698
2880	0.705218	0.863922	0.820180	0.815591	0.617567	0.850709
3240	0.711690	0.859192	0.833919	0.820135	0.643794	0.851277
3600	0.711022	0.857954	0.838290	0.821672	0.661155	0.850901

This experiment used six different feature selection methods. Table 6 is the value of $Mac\mathrm{F}_1$ in different features dimensions for six feature selections. Table 7 shows the comparison of $Mac\mathrm{F}_1$ by using paired T test.

Table 7 shows that the comparisons of $Mac\mathrm{F}_1$ by CHI *vs.* CDDFS do not reject the null hypothesis at a significance level of 0.05, thus demonstrating that using CDDFS cannot improve the classification performance with SVM. But the comparisons of $Mac\mathrm{F}_1$ by OR *vs.* CDDFS, MI *vs.* CDDFS, IG *vs.* CDDFS, DF *vs.* CDDFS seem to reject the null hypothesis, thus demonstrating that using CDDFS can improve the classification performance when using SVM as classifer.

Table 7: Comparison of $Mac\mathrm{F}_1$ by using paired t test

	t	*p*
OR *vs.* CDDFS	-6.841	.000
CHI *vs.* CDDFS	1.864	.095
MI *vs.* CDDFS	-4.714	.001
IG *vs.* CDDFS	-7.971	.000
DF *vs.* CDDFS	-10.658	.000

To sum up, the CDDFS can achieve the best effect and stability of text classification in the majority of tested cases. It demonstrates that the distinguishability and representativeness of feature t for category C can exert an influence on the performance of classification.

4 Conclusions

In this paper, we analyze the distinguishability and representativeness of feature t for category C. And then, we propose the degree of membership and degree of non-membership. At last, we propose the text feature selection method based on category-distribution divergence.

The new method is tested separately with five feature selection algorithms and three classifiers using Sogou corpora. The results show that CDDFS clearly offers a better performance than other five feature selection methods in the majority of tested cases.

However, there are some limitations in our research. Firstly, the experiments only used the Sogou corpora. The other corpora like Reuters or 20 Newsgroups are not being used. Secondly, the experiments only verified the effectiveness in feature selection except the analysis of mathematics principles.

In future work, we would like to use other corpora to test the effectiveness of CDDFS and focus on the analysis of mathematics principles of this method.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No. 71373291).

References

[1] Banka H, Dara S. A Hamming distance based binary particle swarm optimization (HDBPSO) algorithm for high dimensional feature selection, categorization and validation. Pattern Recognition Letters. 2015; 52: 94-100. `http://dx.doi.org/10.1016/j.patrec.2014.10.007`

[2] He S, Chen H, *et al*. Robust twin boosting for feature selection from high-dimensional omics data with label noise. Information Sciences. 2015; 291: 1-18. `http://dx.doi.org/10.1016/j.ins.2014.08.048`

[3] Sardana M, Agrawal RK, *et al*. An incremental feature selection approach based on scatter matrices for categorization of cancer microarray data. International Journal of Computer Mathematics. 2015; 92(2): 277-95. `http://dx.doi.org/10.1080/00207160.2014.905680`

[4] SHEN H, LU B, *et al*. Comparison and Improvments of Feature Extraction Methods for Text Categorization. Computer Simulation. 2006; 23: 222-4.

[5] Li-Qing Q, Ru-Yi Z, *et al*. An Extensive Empirical Study of Fea-

ture Selection for Text Categorization. Computer and Information Science in 2008, Inst. of Elec. and Elec. Eng. Computer Society, United States; 2008. p. 312-5.

[6] Liu H, Chen Q, *et al*. Improved mutual information method of feature selection in text categorization. Computer Engineering and Applications. 2012; 48(25): 1-4.

[7] Wei T, Lu YH, *et al*. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications. 2015; 42(4): 2264-75. http://dx.doi.org/10.1016/j.eswa.2014.10.023

[8] Guo Y, Liu X. Study on information gain-based feature selection in Chinese text categorization. Computer Engineering and Applications. 2012; 48(27): 119-22, 127.

[9] Liu T, Hu L, *et al*. A fast approach for detection of erythemato-squamous diseases based on extreme learning machine with maximum relevance minimum redundancy feature selection. International Journal of Systems Science. 2015; 46(5): 919-31. http://dx.doi.org/10.1080/00207721.2013.801096

[10] Pinheiro RHW, Cavalcanti GDC, *et al*. Data-driven global-ranking local feature selection methods for text categorization. Expert Systems with Applications. 2015; 42(4): 1941-9. http://dx.doi.org/10.1016/j.eswa.2014.10.011

[11] Lee J, Kim D. Mutual Information-based multi-label feature selection using interaction information. Expert Systems with Applications. 2015; 42(4): 2013-25. http://dx.doi.org/10.1016/j.eswa.2014.09.063

[12] Gowid S, Dixon R, *et al*. A novel robust automated FFT-based segmentation and features selection algorithm for acoustic emission condition based monitoring systems. Applied Acoustic. 2015; 88: 66-74. http://dx.doi.org/10.1016/j.apacoust.2014.08.007

[13] Zhang Y, Gong D, *et al*. Feature selection algorithm based on bare bones particle swarm optimization. Neurocomputing. 2015; 148: 150-7. http://dx.doi.org/10.1016/j.neucom.2012.09.049

[14] Zhu P, Zuo W, *et al*. Unsupervised feature selection by regularized self-representation. Pattern Recognition. 2015; 48(2): 438-46. http://dx.doi.org/10.1016/j.patcog.2014.08.006

[15] Lu YH, Cao LC. Text Feature Selection Method Based on Particle Swarm Optimization. New Technology of Library and Information Service. 2011; 208(7): 76-81.

[16] Ghamisi P, Benediktsson JA. Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization. IEEE Geoscience and Remote Sensing Letters. 2015; 12(2): 309-13. http://dx.doi.org/10.1109/LGRS.2014.2337320

[17] LIN D, BAI Q, *et al*. A Ontology-based Document Feature Extraction. Computer Science. 2008; 35: 152-4.

[18] Xu Y, Li J, *et al*. A category resolve power-based feature selection method. Journal of Software. 2008; 19(1): 82-9. http://dx.doi.org/10.3724/SP.J.1001.2008.00082

[19] Wang B. Related Technologies Research on Feature Selection for Text Categorization. Hunan: National University of Defense Technology; 2009.

[20] Sogou Lab. Text Categorization. Available from: http://www.sogou.com/labs/dl/c.html/ (accessed 25 November 2012).

[21] Apache. Apache Lucene. Available from: http://lucene.apache.org/core/mirrors-core-latest-redir.html/ (accessed 26 February 2014) .

[22] Turtle HR, Croft WB. A comparison of text retrieval models. Computer Journal. 1992; 35(3): 279-90. http://dx.doi.org/10.1093/comjnl/35.3.279