Full-text Retrieval Model Based on Term Frequency and Position Weighting

ZhangRui[1,a,*], XiePuzhao[2,b], SunRui[2,c], YangLuchang[1,d] and JiangFeiyue[2,e]

[1]College of Software, Jilin University, Changchun, China

[2] College of Software, Jilin University, Changchun, China

[a] 1364718839@qq.com, [b] xiepuzhao@sina.com, [c] 3250821380@qq.com, [d] 244492644@qq.com,

[e] 1435338725@qq.com

*ZhangRui

Keywords: Lucene, Full-text search, Word frequency position weighting, Computer science, Multi-factor influenc emodel, Simulated annealing.

Abstract. Nowadays mainstream literature retrieval system is based on the search terms, by extracting the document title, keyword, summary of literature to accomplish the function of retrieval. In this article, a full-text retrieval model based on lucene in computer science is purposed. The word frequency weighted algorithm is adopted to set the weighting coefficients in fields of the documents. The computer science literature's attributes are introduced into the evaluation model as an important indicator of the value of literature. The multifactor influence model employs simulated annealing algorithm to fit the best weight coefficients of each factor, making up the defect that Lucene default retrieval method can only retrieve by word frequency. The experimental data were divided into training set and the test set,whose emements are from CNKI.Weights of each field are trained by carrying out feature extraction. Then the model is validated by the test set consisting of a fixed number of high-quality document and inferior ones. The experimental results show that the trained model has higher precision in selecting high-quality documents.

1. Introduction

Nowadays, computer science information emerged at an explosive rate making a large number of workers engaged in IT industry at a loss. However, universities and institutes generally use wide-field literature retrieval model. The information retrieval research aiming for Computer science is scarce. It is important for information retrieval system to meet the growing demands of users.

Contemporary retrieval system evolved from the classical information retrieval model——Boolean model, vector space model[1] and classical probability model[2]. The vector space model proposed by Gerard Salton can rank the query results by similarity and control the output. Robertson and Sparck Jones proposed probabilistic models which is ranking by the descending order of the relative probability, but the documents need to be divided into related and irrelevant sets.

Li Yongchun improved the efficiency of traditional full-text retrieval methods by constructing a Lucene-based full-text retrieval system model in [3]. The system, however, focuses on shortening the retrieval response time but not performance on the accuracy of the results.

In [4] RAM Pereira proposes HTML tag classification, creates linear and nonlinear models, and attempts to determine the optimal weight coefficients for each class of tags. However, the parameters of its nonlinear model need to be artificially determined. In [5], Zhang Chunqing defined the weighted term frequency model—WTFM whose weight coefficients were determined by simulated annealing algorithm. Author ignores the contribution of text content to the results of ranking. Wang Xiangyang proposed a PageRank algorithm based on Citation Network in [6]. Xue Ruiqing and Hu Yifang discussed some specific document attributes which could reflect on its quality in [7.8]. In [10], SP Khapre discussed the theoretical paradigm of information science and computer science in detail, and constructs the theoretical framework of information retrieval.

This model extracts document feature and takes the document properties into consideration by simulated annealing algorithm[11]. The optimal weight coefficients are determined after continuous iteration, The test set, high-quality documents, are used to verify the model’s reliability.

2. Theoretical background

2.1. Classic information retrieval model

The composition of the information retrieval model:

(1) user demand representation: the obtain and representation of user's query.

(2) document representation: content identification and representation.

(3) Matching mechanism: the query mechanism between user requirements and document representation

(4) feedback correction: the search results to optimize

Abstract, the information retrieval model is composed of four tuples [D,Q,F,R(q_i,d_j)]

D: document collection.

Q: user query, the expression of the user task.

F: document representation, query representation and the relationship between them—model frame.

R(q_i,d_j): a ranking function, the function output real number related to the query and document by the value of the function. It can return the ranking list to the user.

The Boolean model is based on the set theory and Boolean algebra. Each of its index words has only two states in a document and the corresponding weight is 0 or 1. The user constructs the query in Boolean logical and submits to model. The search engine determines the query result by pre-established inverted list file.

The vector space model adopts the "partial matching" search strategy instead of exact match of the Boolean model. The basic idea is to use vector to represent text. documents and user queries are expressed as a vector containing the weight of the feature.

The probabilistic retrieval model probes the correlation between the document vector and the query vector, and solves the problem of information retrieval under the probability theory.

2.2. Lucene scoring algorithm

For a given query, a variable is defined to describe the degree of matching between the query and the document. The position of each document in the returned list is determined by this variable.

Lucene combines the Boolean information retrieval model with the vector spatial information model, and the scoring formula is:

$$\text{score(q,d)=coord(q,d)*queryBoost(q)*}\frac{\text{V(q)*V(d)}}{\text{V(q)}}\text{*lengthNorm(d)*docBoost(d)} \qquad (1)$$

In practice, formula(1) can be translated into

$$\text{score(q,d)=coord(q,d)*queryBoost(q)*}\sum_{i\in q}\text{(tf(t}\in\text{d)*idf(t)}^2\text{*boost(t)*norm(t,d))} \qquad (2)$$

coord(q,d): Scoring factor, related to the query items which appear in the document.

queryNorm(q): The effect of the query weight on the score.

tf(t϶d): Term t, The number of occurrences in document d.

idf(t): Inverse Document Frequency.
boost(t): The weight of the query term.
norm(t,d): The weighting factor of the length of field.

3. Attributes value evaluation index

3.1. Authority evaluation index

The value function $Au(i)$ of the authority of periodicals and authors is defined as follows:

$$\mathrm{Au(i)}=\alpha*\frac{\mathrm{Nf_i}}{\sum_{j=1}^{N}\mathrm{f_i}}+\beta*\frac{\mathrm{Np_i}}{\sum_{j=1}^{N}\mathrm{p_j}} \tag{3}$$

In the formula(3), f_i is an influence factor of the journal i. p_i is the total number of documents published by the authors of the literature i. α, β are the weights of periodical authority and author authority respectively, indicating the influence degree of each factor on the value of the document. N is the number of documents involved in the literature ranking, and the higher the authority function value, the greater value of the documents and such essays are more likely to meet the demand of the scientific researchers.

3.2. Publishing time evaluation index

A great deal of data analysis shows that innovations in computer science and computer moves forward at a staggering rate. The value of the literature declines with age. When the impact of publishing time on the value of computer science literature is considered. within normal errors, function T(i) is defined as follows:

$$\mathrm{T(i)=timeNow-timePub(i)+b} \tag{4}$$

timeNow refers to the current user’s query time. timePub(i) refers to the publishing time of one specific document. b is a constant term used to improve the fitting degree .According to the survey data, for computer science research, the Time function T(i) is be approximated as linear function. Up-to-date documents in computer science tend to have higher function value.

3.3. Citation evaluation index

Citation frequency is an important criterion to evaluate literatures values. Although up-to-date papers are more innovative and potentially valuable, it is certainly that they are not cited as many times as the older ones. To reasonably evaluate the citations’ reflect on literatures value, a hierarchical evaluation index Fre(i) based on the cited times of literatures is purposed as follows:

$$\mathrm{Fre(i)}=\begin{cases}0,i=0\\1,0<i\le4\\2,4<i\le8\\4,8<i\le16\\8,16<i\le25\\16,25<i\end{cases} \tag{5}$$

The most cited ones are milestones or that have pioneer effect in their fields. Although this kind of literatures' significance are unneglectable. They are lack of reference value as the continuous summary by scholars. When a document’s citations is more than 25, it can be assumed that the document is not an up-to-date. There is no need to continue grading by the cited time. When the cited times is between (0,4]，this function sets a relatively small evaluation score, roughly excluding the effect of author's own reference to his literature on the ranking of returned list. If the citation of a document is 0, the document almost has no reference value, or it is just published very recently. In either case, the literature system will not tend to return such documents.

4. Full-text retrieval model based on term frequency and position weighting

4.1. Literature quality evaluation criteria

Lucene-based term frequency retrieval is improved in this article. The search scope is expanded from three originally to four domains: title, keyword, abstract and content. Weighting coefficients are introduced for each domain.

Different fields of a scientific literature differ in the degree of generalization to the literature. Title can clearly indicate the subject and object of the research. Keywords can be used to summarize the technical implementation involved. Abstract could briefly describe the purpose, method, and result of the research. Literature content is the expansion of specific operations. It is not reasonable to perform an identical process for the match between the four parts and the search word. And different weight coefficients w_i for each part are introduced. Formula (2) is improved:

$$s\_score(q,d)=\sum_{i=1}^{N} w_i*score_i(q,d) \tag{6}$$

s_score is the sum of weighted individual fields $score(q,d)$. w_i is the weight of the i field of a document (title, key words, abstract, content). $score_i(q,d)$ is a matching score between the user query to the i field of the document. N is the count of fields in one specific literature.

The relative value of the weights represents the contribution of each domain to the ranking of the literature in terms of text relevance.

In addition to the matching degree between the query term set and the documentation set, the author authority, Impact Factor of Academic Journal, the publishing time and the number of citations are all factors that determine the value of a document. Formula (7) can approximately show the attributes value of the literature.

$$v_{score(q,d)}=Au(i)+\varphi*T(i)+\mu*Fre(i) \tag{7}$$

φ and μ are weights coefficient for the publishing time evaluation index and the citation evaluation index.

Different query terms have different matching degree scores s_score for same literature, but the v_score of an article is fixed. Its value is determined at the moment the document is published.

By the query—text matching score and document attribute score, the value of an article to the user's demand can be finally determined as formula (8).

$$u\_score(q,d)=s\_score(q,d)+v\_score(q,d) \tag{8}$$

The Full-text retrieval model based on term frequency and position weighting—FTFW is defined.

4.2. Establishment of weighting coefficient

In the FTFW model, weight coefficients are needed to be determined. Results could be extremely bad if weight are designated from experience. However, it's easier to collect a list of High-quality documents（experts ranking list） for given query terms. This article provides a method based on simulated annealing algorithm, combining with the expert ranking list to extract features from the relationship between the given keywords，content and the values of attributes for determining the appropriate weighting coefficients. The determination of the weight coefficients is transformed into an optimization problem.

The outline of this model:

Construct the penalty function, compare the expert List Ranking with the calculated List Ranking, find a proper algorithm to represents the distance between them, search in the weight combination neighborhood iteratively that could make the penalty function value smaller. The model uses simulated annealing algorithm to calculate the minimum distance between the calculating ranking list and the experts ranking list. In the face of the new query, the best documents can be extracted from the set of corresponding documents.

Simulated annealing algorithm shown in Figure 1:

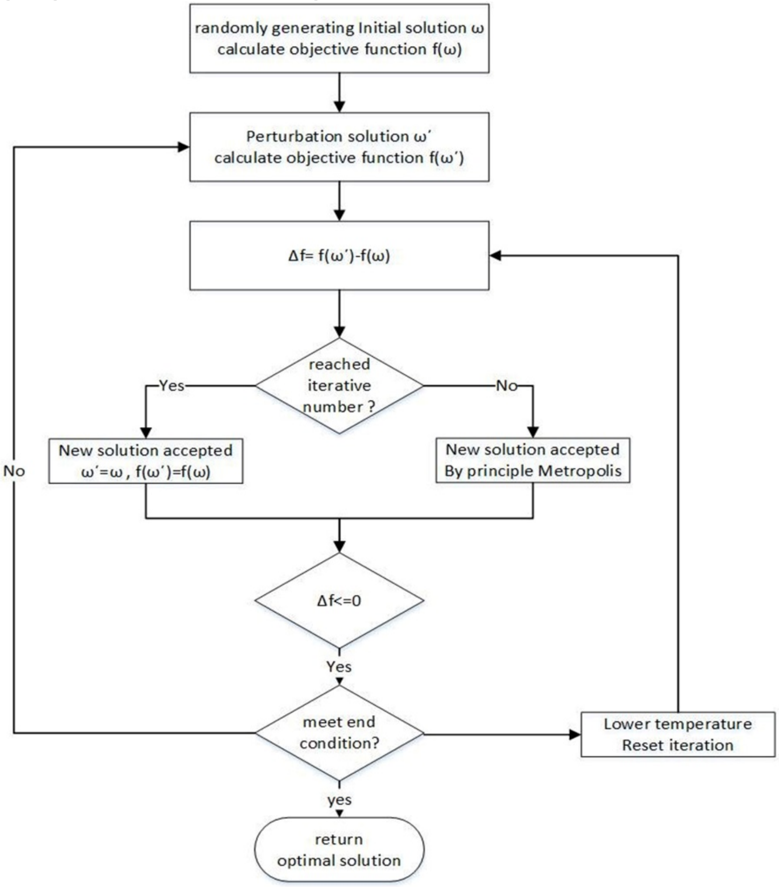

Figure 1 Flow chart of simulated annealing algorithm.

5. Experiment and analysis

5.1. Experimental design

FTFW, a lucene-based model, is implemented in the Java language. The data experimental data – 200 documents in computer science from CNKI is divided into two groups (100 in each group) by the query term. The first group is the training set, all made up of high-quality documents. An expert list is constructed on the basis of the quality of documents. The second group is the test set, which consists of 20 high-quality literatures and 80 relatively inferior ones selected by manual screening and then rearrange them into sequences.

After weight coefficients are trained by training set, the model scores each document in test set, and then select twenty documents from top to bottom in this calculated ranking list. The precision of the model is calculated by comparing the selected 20 documents with that in the expert ranking list.

A reasonable evaluation mechanism is needed to judge the rationality of the current weight combination. Formula (9) is introduced to measure the distance between a calculated ranking list and an expert ranking list

$$D\left(R,R^{'}\right)=\frac{\sum_{i=1}^{n}(n-i)*|j-i|}{\sum_{i=1}^{\left\lfloor\frac{n}{2}\right\rfloor}[(n-i)*i]+\sum_{i=\left\lfloor\frac{n}{2}\right\rfloor+1}^{n}\left[(n-i)^{2}\right]}\left(R_i=R_i^{'}\right) \qquad (9)$$

n is the total number of articles involved in the ranking list. R_i represents the i article in the calculated ranking list. $R_j^{'}$ represents the j article in the expert ranking list. The value of $D(R,R^{'})$ reflects the difference between the expert ranking list and the calculated ranking list. The number and position of the misplaced documents related to the cost is taken into consideration. The numerator of formula (9) is used to measure the actual distance between the two ranking lists，The denominator is used to normalize the outcome. For this evaluation mechanism, the cost of the misplaced literature at the very top is far greater than that at the bottom, because the purpose of the model is to pick out high-quality documents from the varied quality literatures.

5.2. Experimental analysis

The weights of the model are trained by the simulated annealing algorithm by the data of group 1. The convergence rate of the distance between two lists is shown in (Figure 2). As shown in (Table 1), when the iterations are approximately in range from 2674 to 2737, the lists distance approaches convergence, the value is 5.23396. The weight coefficient trained at that time can be approximated as the optimum weight coefficient (shown in Table 2)

.

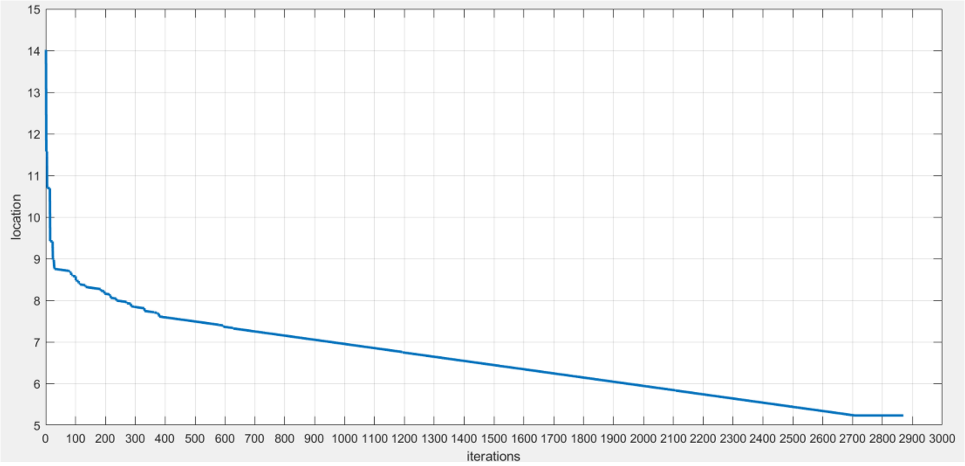

Figure 2 The relation between iteration and distance.

Table 1 The relation between iteration and distance.

iterations	1	2	6	17	104	278
distance	14.0157	11.5804	10.6985	9.42092	8.4664	7.9248

iterations	1168	1963	2421	2674	2705	2737
distance	6.78384	5.981	5.51996	5.23396	5.23396	5.23396

Table 2 Weighting coefficient.

w_1	242.8731273
w_2	257.4028018
w_3	264.4352781
w_4	235.928349
α	15.28813698
β	250.1290051
φ	-52.0845013
μ	116.4434305

Score and rank documents in group 2 with the best combination of weights. There are 20 high-quality documents in the group 2, so the top 20 of in the calculated rank list are selected. Formula (10) calculates precision rates for estimating model performance by analyzing the selected documents and high-quality documents in the test set.

$$\text{precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{10}$$

TP is the number of high-quality documents and FP is the number of inferior ones selected from the calculated ranking list. Retrieval system's ability of selecting high-quality documents can be evaluated by Formula (10). By experiment, the precision of the model is: 0.65. To test whether the FTFW model is improved compared to the default Lucene model, next experiment employs Lucene default ranking algorithm to test the same document set with the same query, and its accuracy is: 0.30. It's shown that the precision of FTFW model is 0.35 higher than the default algorithm

5.3. Discussion of experimental results

The experimental result shows that the reliability of FTFW model is superior to Lucene's default ranking algorithm in computer science. This model can pick out high-quality documents, however the ability of accurate ranking is not prominent. The reason may be that the attributes of the document are not independent of each other, and there are varied degrees of connection between different attributes. This system does not construct semantic net. In fact, different expressions of the same retrieval words should not cause a great change on the ranking documents lists.

6. Summary and future work

The main Tasks are described as following:

(1) Each field of the document are weighted on the basis of Lucene full-text retrieval system.

(2) The literature's attributes including the author authority, Impact Factor of Academic Journal, the publishing time and the number of citations are introduced into the evaluation model as an important indicator of the value of literature

(3) the features of computer domain literatures are extracted by employing simulated annealing algorithm, and the weight of various attributes affecting the quality of the document is determined.

(4) The matching precision of the calculated list and the expert list is used as indicator. Through the designed experiment and the feedback of comparison between the FTFW model and the Lucene default model, it is found that the precision of the former model is higher.

Future work: try to build semantic net, through the expansion of the semantic network, a more accurate literature evaluation model will be established. The citation network combined with PageRank algorithm and analysis of the relationship between citations could be used to establish a more optimized model.

References

[1] Salton G, Lesk M E. Computer Evaluation of Indexing and Text Processing[J]. Journal of the Acm, 1968, 15(1):8-36.

[2] Robertson S E, Jones K S. Relevance weighting of search terms[J]. Journal of the Association for Information Science & Technology, 1976, 27(3):129–146.

[3] LiYongchun,DingHuafu. Lucene's full-text search research and application [J]. Computer Technology and Development | Comput Technol Dev, 2010, 20(2):12-15.

[4] Pereira R A M, Molinari A, Pasi G. Contextual weighted representations and indexing models for the retrieval of HTML documents[J]. Soft Computing, 2005, 9(7):481-492.

[5] ZhangChunqing, ChenChao, ShaoZhengrong. Similarity evaluation model of information retrieval based on weighted word frequency [J]. Computer Simulation, 2008, 25(1):134-137.

[6] WangXiangyang,MaJun. A PageRank-based scientific literature quality evaluation algorithm[J]. JournalofGuangxiNormalUniversity(NaturalScienceEdition), 2009, 27(1):165-168.

[7] XueRuiqing. Research on Ranking Prediction Algorithm Based on Author 's Authoritative Value[D]. Jilin University,2012.

[8] HuYifang. The relationship between the citation of academic papers and academic quality[J]. Library Tribune | Lib Trib, 2015(5):56-59.

[9] YaoZhichang, DengQun, LiChengjun,. Reflection on the Value Evaluation of Scientific Papers [J]. Chinese Journal of Scientific and Technica, 2005, 16(6):795-797.

[10] Khapre S P, Basha M S S. A Theoretical Paradigm of Information Retrieval in Information Science and Computer Science[J]. International Journal of Computer Science Issues, 2012, 9(5):232-240.

[11] Bandyopadhyay S, Saha S, Maulik U, et al. A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA[J]. IEEE Transactions on Evolutionary Computation, 2008, 12(3):269-283.