International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 3– No.1, July 2012 – www.ijais.org

Social Business Intelligence Text Search System

Sagar Ligade
ME Computer Engineering.
Pune Institute of Computer Technology
Pune, India

Pravin Game
ME Computer Engineering
Pune Institute of Computer Technology
Pune, India

ABSTRACT

Today the search engine plays the important role in the internet community. The search engines are in various forms such as social search, personalized search and enterprise search. For satisfying the end user, developers are developing new ranking algorithms to satisfy the user need. In proposed system we use the vector space model merged with the social search to provide new ranking algorithm to match query with the document. We propose new search system architecture to search the document by index terms.

Keywords

Information Retrieval and Indexing, Query Intent, Social Search.

1. INTRODUCTION

There are three classic information models present today. They are Boolean, Vector and Probabilistic model[1].The classic information models in the information retrieval consider that each document is presented by the index term k_i and document is represented by d and $w_{i,j}>0$ be a weight associated with the pair(k_i,d_j).

Boolean model: In Boolean model the index term weight variables are all binary i.e. $w_{i,j}$ can take either 0 or 1 value. The Similarity of the document d_j to the query term q is defined as Sim(d_j,q)=1, if term is present in the document otherwise it is 0.

Vector model: The Vector model partially match the query terms with the documents. In this model we get the ranked document result set. The similarity of the document d_j to the query q is defined by the cosine of angle between them[2].The apache lucene[3] uses the vector model for information retrieval.

Probabilistic model: In this model we are specifying the properties of an ideal answer set. User takes a look at the retrieved documents and decides which ones are relevant documents.

These various models are getting modified in the recent year. User can provide feedback on the content of search results and we use this information to re-rank the documents. In the recent days developers used social factors to rank documents[4][5].

We proposed new retrieval strategy which is based on the vector space model and for the retrieval. We propose the new architecture for retrieval of the information which gives the social weight to the index term of the document.

2. RELATED WORK

Social Search:

Social Search is the type of web search “web 2”.Social web is increasing rapidly from its appearance. It guides user for creating the correct query and helps the user for what they need. Search results produced by social search give more visibility to content created or touched by users. Zhang et al.[6] models the large scale social network on live online community and evaluate user relationship index from the model. Aardvark is the social search engine which finds the right person to satisfy a user information need [7].

3. SOCIAL RANKING APPROACH

As we know most of the web search engines focus on the text and the document similarity. We introduce the social ranking for the document and this social ranking decides the relevance between query term and the document.

The social ranking of the documents are based on the user social graph. We make the use of social networking services to rank our search document.

The comments on search documents, ‘likes’ and the recommendation for the URL are considered to be the user gestures. System uses this user generated content to provide clues for user to get information.

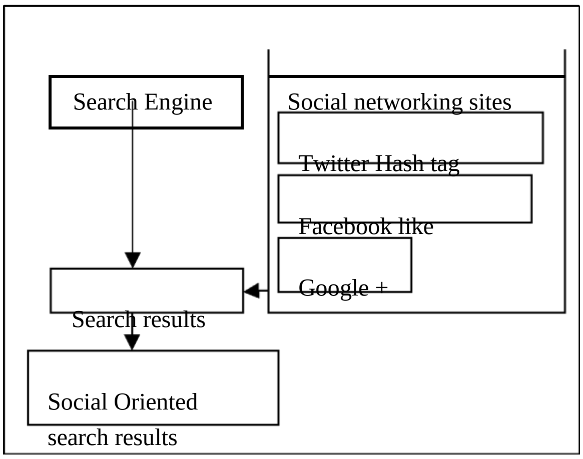


Fig 1: Social ranking of the search results.

The social rank of the document is calculated from the user gestures. 'Likes' of users' friend on particular document or comment on the document can be considered as user gestures for the document.

The social indicator of the documents is captured so that system knows that content is liked by the previous user[8]. The social networking sites provide this facility.The social rank of the given term can be calculated from the previous user generated data [9].There are other user gestures such as the sharing the documents and this information can be used to built the social rank of the document.

4. SYSTEM ARCHITECTURE.

System provides the search interface to the end user. System modifies the traditional search engine by adding index term for search to documents which taken from the social graph of the user. The Index term calculation is done by the luke toolbox [10].

The Luke toolbox gives distinct keyword information about the document. We used this information for searching document. The Lucene Connector uses the 'Document' methodology to index the My-SQL database.

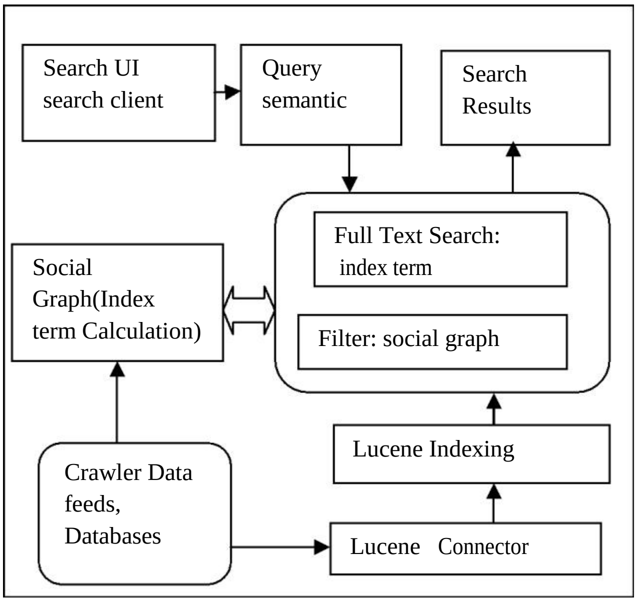


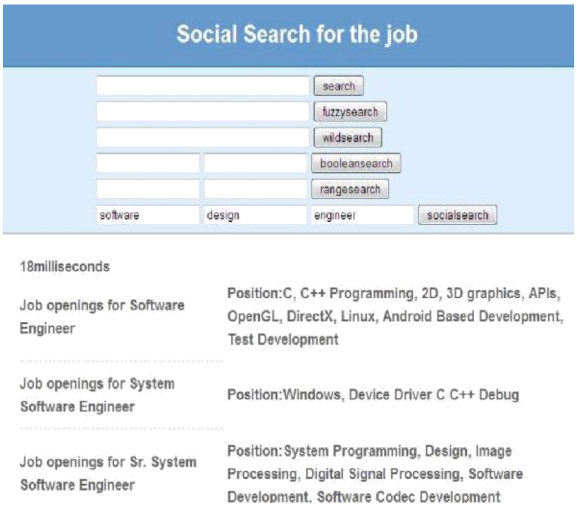
Fig 2: The system architecture

The document indexing is done with the help of index writer of the lucene.

The luke toolbox provides the distinct keywords and list of most frequent terms. In our implementation we used the luke toolbox to find the most frequent terms in the document and in future we will retrieve this frequent terms from the trusted user friend from social graph of the user for the document indexing.

User Interface:

The user interface of the application is designed which is similar to most of the web search engines. The user can interact with system with the help of luke toolbox for searching the documents. The user enters the most important keyword in the social search. Query formulation is done with the help of luke toolbox. From the luke toolbox user comes to know most frequent keywords in the document so user analyzes the trends in the searching of documents.



software design engineer

18milliseconds

Job openings for Software Engineer	Position:C, C++ Programming, 2D, 3D graphics, APIs, OpenGL, DirectX, Linux, Android Based Development, Test Development
Job openings for System Software Engineer	Position:Windows, Device Driver C C++ Debug
Job openings for Sr. System Software Engineer	Position:System Programming, Design, Image Processing, Digital Signal Processing, Software Development. Software Codec Development

Fig 3: Integrated user Interface with the help of luke toolbox.

We used the separable search for the document retrieval in which some hints for searching the documents are provided. For example in fuzzy search even if typographical error occurs in the index term still system searches the document. In wild search the initial letters of the index term are required for searching the documents and in the range search we give the range limits, in that limit all the search results are displayed.

5. EVOLUTIONARY ALGORITHM APPROCH FOR THE SOCIAL SEARCH

An EA uses some mechanisms inspired by biological evolution:selection,reproduction,mutation,crossover[11]. We can use the same approach for the information retrieval. In paper [12][13] the genetic algorithm is used in which the fitness function is based on the cosine similarity.

The main steps are illustrated as below.

1. Selection:

In selection process, we represent each document as in lucene index form. These lucene index forms are called the initial population that will feed into genetic operator process. The selection of the particular document depends on its fitness value. System introduces the new fitness function for the selection. The fitness value depends on the Vector space model score and the social score.

The constraints on the fitness function are user gestures on the particular document i.e 'likes', commenting on the results, sharing the document in social network.

2. Reproduction: It consists of the two genetic operators

A. Crossover

It is the genetic operator that mixes two chromosomes together to form new offspring(Search Documents).In the Social Search system the SearchManager calls different methods and creates the Crossover between the SearchResultBean.

B. Mutation

It involves the modification in values of each gene of the solution with mutation probability. In the Social Search system boostsearch can modifies the gene value of the query string.

3. Replacement:

The initial population is updated by replacing some existing solutions by the newly created ones. The SearchResultBean can be used for the replacing the search results.

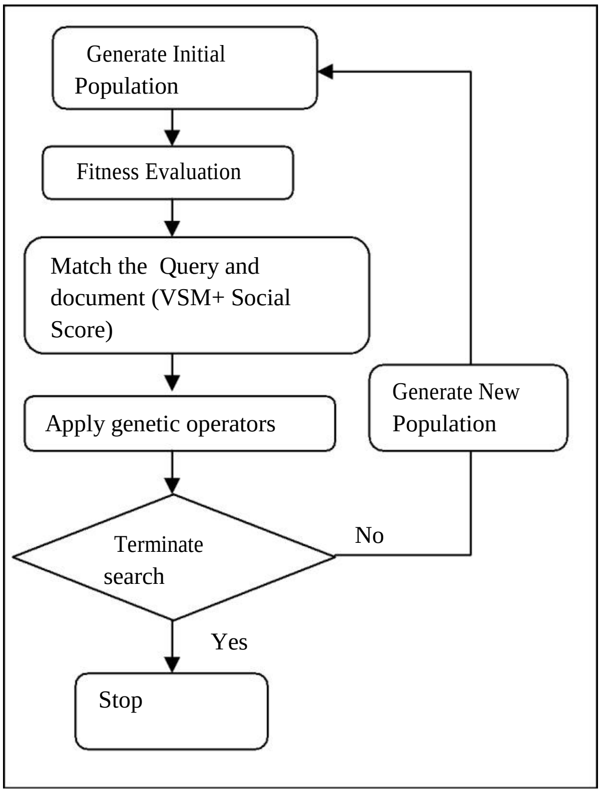


Fig 4: Genetic model for the Information retrieval.

Query chromosomes are formed from the user query. The document chromosomes are the keywords inside the document itself. We used TF-IDF[2] methodology for the searching mechanism.

6. SYSTEM ARCHITECTURE IN THE FORM OF MATHEMATICAL MODEL:

The proposed system ‘S’ is defined by the set theory and the problem is described:

S={q,d,$w_{i,j}$,$w_{i,q}$,t,N,k,a,b,c|Sim(d_j,q),SocialBoost(d),DCG}

q=query to the system.

d=document set.

$w_{i,j}$= weight associated with pair(k_i,d_j).

$w_{i,q}$= weight associated with pair(k_i,q).

t=total no of index term.

N=total number of index term in the document.

k=Generic index term.

Sim(d_j,q)=cosine similarity+SocialBoost(d)……..(1)

Sim(d_j,q) is based on cosine similarity[2].

$$\mathrm{Sim}(d_j,q)=\frac{\sum_{t=1}^{t} Wi,j\times Wi.q}{\sqrt{\sum_{i=1}^{t}(Wi,j)^2}\times\sqrt{\sum_{j=1}^{t}(Wi,q)^2}}\ldots\ldots\ldots(2)$$

Where,

$$Wi,j = fi,j \times log\frac{N}{ni}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

The inverse document frequency can be calculated by the following formula, The equation (3) is derived from the equation (4) and (5).

$$idf = log\frac{N}{ni}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Also $f_{i,j}$ can be calculated as,

$$fi,j = \frac{Freqi,j}{maxlfreqi,j}\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

Freqi,j= The number of times the term k_i is mentioned in the text of the document d_j

n_i = be the number of documents in which the index term k_i appears

Social Boost(d)=Social boost of the document.

The Social rank of the document is calculated by the linear scaling method.

Social Boost(d)=a+ b+ c

a, b, c,=User gestures for the search results.

a=Facebook likes(index term)

b=twitter commenting on the results (hast tags)

c=Google+ recommendation (index term)

Discounted Cumulative Gain[14] is the measure of the effectiveness of the document retrieval.

Input to the System:

I/P=(q,d,a,b,c).

In our system, the problem is to maximize the probability of relevant document which are selected from the document.

When the social boosting on the document is done then system displays the search results according to the index term.

7. EXPERIMENTAL RESULTS

The Discounted cumulative gain (DCG)[14] is a measure of effectiveness of the document retrieval. We make the graded relevance scale for the document which are obtained in response to a search query, an experiment participant is asked to judge the relevance of each document to the query. Each document is to be judged on a scale of 0-3,0 meaning irrelevant, 3 meaning completely relevant, and 1 and 2 meaning "somewhere in between" ,for the documents ordered by the ranking algorithm. We gathered the documents related to the job information from various job information sites. System indexed the documents with the help of lucene.

Relevance grading for the search results are calculated as follows

Table 1: The graded relevance from the end users

List of document	User1	User2	User3	User4	User5
D 1	3	3	3	3	3
D 2	3	3	2	1	1
D 3	2	2	2	2	2
D 4	3	1	1	0	3
D 5	1	1	0	3	1

Table 2: The DCG measure for the user 1.

Document i	Relevance i	Log_2 i	Relevance/log_2 i
D1	3	0	N/A
D2	3	1	3
D3	2	1.59	1.25
D4	3	2	1.5
D5	1	2.32	0.4310

DCG measure for user1 =3+3+1.25+1.5+0.431=9.181
The average relevance is 2.4
The Average DCG measure for user1 = 3+3+1.5=7.5

DCG measure for user2 =3+3+1.25+.5+0.4310=8.181
The average relevance is 2
The Average DCG measure for user2 =3+3+1.25=7.25

DCG measure for user3 =3+2+1.25+.5=6.75
The average relevance is 1.6
The Average DCG measure for user3 =3+2+1.25=6.25

DCG measure for user4 =3+1+1.25+1.29=6.54
The average relevance is 1.8
The Average DCG measure for user4 =3+1.25+1.29=5.54

DCG measure for user5 =3+1+1.25+1.5+0.4310=7.181
The average relevance 2
The Average DCG measure for user 5=3+1.25+1.5=5.75

The graph is as follows:

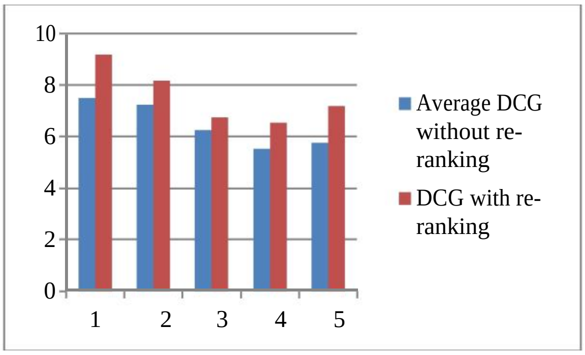


Fig 5: DCG measures

On X-axis we get the user1,2,3,4,5 and on Y-Axis we will get the DCG measures(0-10).

We analyzed our System in Three dimension in which we represent x axis with the query term i.e. input is given to the system and on Y axis we got the documents as a response to the user query .We set the performance parameter i.e.DCG measure on the Z axis.

The graph drawn as follows:

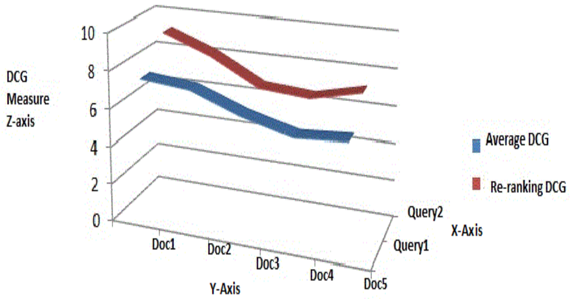


Fig 6 : Improved DCG by using social index term. The graph shows that the DCG after re-ranking of the search documents are more than the average DCG value.

8. CONCLUSION

Proposed Business intelligence text search system uses the vector space model algorithm and adds social indexing for document retrieval. The boosting of the index term according to the user preference is the main task of the work. With the help of user gestures which is collected from social networking sites documents are indexed and searching depends on the social index term. The new modification improves retrieval of the documents and promising search results are obtained from the social search. In the future social term calculation can be made more complex for the document retrieval by giving importance to the trusted users for calculating the social term.

9. REFERENCES

[1] R Baeza-Yates and B ribeiro-Neto. “Modern Information Retrieval”, Pearson education 1999.

[2] S.G Salton, A Wong C.S Yang,A Vector Space model for automatic indexing,Communication of ACM,vol 18n 11,Nov.1975,pp 613-620.

[3] http://lucene.apache.org/java/docs/index.html. (HTML)Retrieved 2011-08-01

[4] Jiandong Cao,Yang Tang,Binbin Lou, “Social Search Engine Research”, IEEE Conference on Computer Science and Information Technology,vol 7, pp.308-309,2010.

[5] Mohammad Ali Ghaderi,Nasser Yazdani Behzad Moshiri, “A Social Network Based Meta Search Engine” IEEE Conference on ISTEL,pp744-749,2010.

[6] Lu Zhang ,Yanlong Wen,Haiwei Zhang,Ying Zhang,Xiaojie Yuan, “User Relationship Index based on Social Network Community Analysis”,IEEE Conference on Business Management and Electronic Information,vol 4,pp.66-69,2011.

[7] Damon Horowitz,Sepandar D.Kamvar, “The Anatomy of a Large-Scale Social Search Engine”. Proceedings of the 10th ACM Conference on World Wide Web, Raleigh, North Carolina, USA,2010,ACM.

[8] Facebook,https://developers.facebook.com/docs/reference/api/ (HTML)Retrived on 2011-09-02.

[9] Brynn M Evans,Ed H.Chi, “Towards a Model Of

Understanding Social Search". Proceedings of the 8th ACM Conference on Computer supported cooperative work,2008,ACM.

[10] http://code.google.com/p/luke/ (HTML)accessed on 2012-04- 04.

[11] Goldberg,David E(1989),"Genetic Algorithm in Search Optimization and Machine Learning",Kluwer Academic Publishers,Boston,MA.

[12] Ahmed A.A, Radwan Bahgat, A.Abdel Latef,Abdel Mgeid A Ali and Osman A.Sadek, "Using Genetic Algorithm to Improve Information Retrieval System" World Academy of Science Engineering and Technology,vol.17,Feb2006.

[13] Siti Nurkhadijah Aishah Ibraim,Ali Selamat,Md Hafiz Selamat, "Optimization of E-business Social Network Mapping Using Genetic Algorithm". International Symposium on Information Technology.vol2.pp 1-7,2008. [14] Kalervo Jarvelin, Jaana Kekalainen: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002).