ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Document Ranking Using Customized Semantic Networks of Text Documents and the User Query

Shubhra Joshi, S. S. Sonawane

Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India

ABSTRACT: As technology is advancing in each area ,effective and efficient retrieval of documents is a basic need.Classic Information Retrieval systems does not only aims at finding the relevant information but they are supposed to retrieve the most relevant documents as well as rank or organize retrieved documents according to its degree of relevancy with the given query.The main task is to decide which documents are relevant and which are not so as to satisfy users' information need.Traditional ranking approaches mostly rely on exact 'bagof-words' strategy to decide the relevancy of the data with the given query. The relevancy score was defined in terms of number of times the words that is in the query appear in the document i.e. term frequency. A semantic search framework could be the solution for this problem which is then adopted in several studies.While retrieving documents for any query contextual meanings of words from documents and queries should be considered to match non-matching terms. Semantic search framework uses such core semantics of data exploiting domain ontology.Here a semantic search framework for semantic pre-processing of documents and queries is adopted which considers relationships among the words in the document also in the query (e.g. is-a, partof, sub-part etc.) with the help of semantic network generation algorithm.The probability of matching users' information needs with stored documents can be increased by considering meaning of terms from data thus by exploiting relationships among terms.

KEYWORDS: Document semantic network, Domain ontology, Engineering document, Personalized search, Ranking, Semantic search, Retrieval, User profile

I. INTRODUCTION

In engineering domain, it is very difficult to improve engineers' productivity on the basis efficiency of document retrieval.In engineering domain, various product development process are executed simultaneously in which huge amount of documents are created and accessed by engineers.Major improvement is required in the existing document retrieval approaches to fulfill engineers' information needs.Most of the traditional methods uses 'bag-of-words' strategy since it is very popular and easy to execute.In this strategy documents are retrieved by keyword-based method by comparing exact matching terms only between document and query.Such approach fails to find most relevant documents in engineering domain,since engineering documents differs from other documents in terms of syntax formulations and core semantic complexities. Most of the engineering based documents contains large number of acronyms and abbreviations, which results in high brevity of documents.These characteristics disables the system to retrieve relevant documents for all users' information needs.Thus to improve retrieval performance of existing systems various techniques can be applied such as document partitioning, query expansion, or document ranking methods.But it is necessary to compare and investigate among all these techniques, to improve classic document ranking methods; domain ontology helps not only to represent the core semantics of document more precisely but to calculate the relevance score of each document.Thus in the proposed study, a domain ontology based new document ranking technique is used.

Semantic search approach represents documents and queries in the form of graphs by exploiting ontology, where each node represents each terms and edges represents relationships among them on the basis of which dependencies between two terms can be recognized.Most of the times, non-matched terms could be the important factor to decide documents'

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

relevancy.Thus, recovery of such semantically related nonmatching words to exact matched words can be done with the help of domain ontology.However, to improve performance of existing approaches a more elaborative and efficient relevancy decision schemes are necessary.One of the major advantage of semantic based approach is that it helps to improve the accuracy of retrieved results for any user query by considering users' intent and the actual meaning of the terms present in the given query so as to output most relevant documents ,which results in the betterment of search engines' performance.Now,one of the important decision is to evaluate semantic similarity between any two terms either from query or document, for that various methods can be used which are further explained in the literature survey.In many approaches, Wordnet is the most widely used reference database for calculation of semantic relationships between terms. In the field of huge semantic network, various algorithm have been proposed for ranking multiple semantic paths between any two nodes, as there is a chance of more than one relationship could be present between those two terms.So, deciding relevance score for each semantic path is an important factor which uses various weighting schemes to find out the accurate relevance score.Hence, these weighting schemes contributes majorly in document various document ranking techniques.

A. Background

Most of the traditional document retrieval methods have some major drawbacks which majorly influence the performance of the relevant document retrieval process are listed below:-

- Bag of words concept never considers semantic relationships between terms, though those terms are semantically related to each other(e.g. has-a, under-this, etc.)
- Relevance score of a document with respect to any query is dominated by number of exactly similar terms between a query and a document completely.

II. REVIEW OF LITERATURE

QE is considered a viable solution, expanding process by expanding query keywords with related terms. AHP is an effective tool for dealing with complex decision making; it aids the decision maker to determine the priorities of used criteria [1].

Learning to Rank is applied to automatically learn a Ranking function from the training data. In this paper, a new hybrid learner is introduced based on NN and SVM that gives better performance than learning using NN alone [2].

Proposed system presents an improved Semantic Similarity technique to rank a web page from a set of given web page which access the user history to rank the webpage according to the user query. An online interface is developed using asp.net web technologies and c.net is being used as a programming language tool [3].

The proposed algorithm calculates page rank value or importance of web pages based on the visits of incoming links on a page. It observed that the page which has more visits of incoming links is carrying more rank value than less Visited pages [4].

The method proposed in this paper uses the concepts and relationship between the concepts that exists both in the document and the user query to improve the retrieval of relevant document. A different method is used for keyword extraction, hence it leads to better results [5].

The proposed paper explores the problem of maintaining the semantic relationship between different plain documents over the related encrypted documents and give the design method to enhance the performance of the semantic search [6]. In this paper, we have proposed a new ranking strategy known as SemRank which uses the SI measure to calculate the image relevancy weights against the query. It has an advantage that it can employ the semantics inside the image and the query in determining the ranking order [7].

III. SYSTEM OVERVIEW

In the proposed work, documents relevance score is calculated using a DSN (Document Semantic Network) generated from a document. After forming DSN for each document ,relevancesocre is measured for that DSN which ultimately decides the ranking of the particular document to user query.

If a DSN for any document contains more number of relationships between terms that are related to users intent, then document will be considered as highly related to that particular query. Various measures are exploited to estimate the weight of each relation against the information needs.Personalized ranking can be also provided to each user by

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

considering its all previous searches,which is nothing but user profile. This complete work can be done by following two steps :

- Generation of document semantic network for a document.
- Relevance score measurement for each document semantic network of a document.

Generally, in all existing models Boolean model matching concept is used for retrieving documents after submission of an query by user. But, after retrieving documents, ranking them in a correct order is a major task to do, for that purpose document semantic networks are generated. DSN's enables ranking the retrieved documents efficiently by considering semantics of the document and the given query. During this process,personalized relevance score for each user is also provided by exploiting user profile.After all this processing, retrieved document result are arranged in descending order of the relevance score.

A. System Architecture

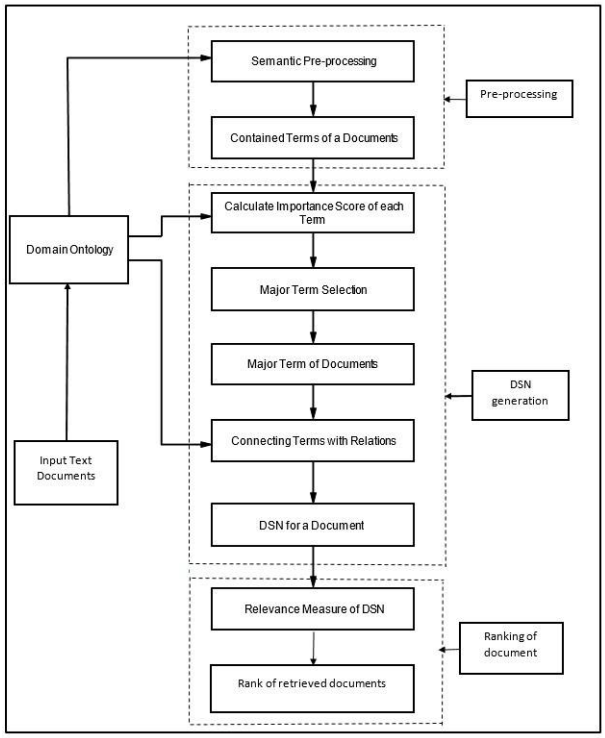


Fig. 1. System Architecture

IV. PROPOSED METHODOLOGY

As shown in system architecture , documents are firstly pre-proccesed in which various tasks such as stopword removal, stemming etc. are performed. Properly pre-proccesed documents also might contains large number of words which makes it difficult to form DSN for each document. Thus, representing the core semantics of document with minimum number of words or terms from the document is an important and difficult task.Thus, following various emasures are used to decide important terms from document to be considered. Importance measurement for terms in a document:To decide which terms are minor or major information elements in the document, term weighting schemes using three measures are introduced. The three measures are

(1) Structural Score

(2) tfidf Score, and (3) Semantic Score

The structural score of term I_{ij},StS (Iij) , is computed as:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

$$StS(I_i^j) = \frac{|Sub(I_{ij})|}{max|Sub(I_{kj})|} \quad (1)$$

Second, the tfidf score for term Iij is calculated using the term frequency and the document frequency. Term frequency (tf) and inverse document frequency (idf) are the foundations of the most popular term weighting scheme in IR. The tfidf score of Iij ,tfidf(Iij), is computed as:

$$Tf{-}Idf(I_i^j) = Log(tf(I_i^j,d_j)+1)*Log(|D|/1+df(I_i^j,D)) \quad (2)$$

Where tf (Iij ,dj) is the frequency of term Iij within document j and df (Iij , D) is the no. of documents that contain term Iij in the document collection D. Thus, terms with a high tf and low df will get high tf-idf scores.
Third, the semantic score of a term Iij, SSr(Iij) is computed as:

$$SS_R(I_i^j) = \frac{|(SD(I_i^j, I_k^j) \leq R)|}{|d_j|, (1 \leq k \leq |d_j|)} \quad (3)$$

where SD(Iij , Ikj) , called the semantic distance , is the minimum no. of hops between Iij and Ikj in the ontology, and R is a parameter to define the semantic distance range.The semantic score measures how each term is far from the core semantics of a document. Through this measure, briefly mentioned terms in a document could be revealed. The importance score function IS (Iij), linear combinations of the three measures, is then calculated as:

$$IS(I_i^j) = w1.StS(I_i^j) + w2.tf - idf(I_i^j) + w3.SS_R(I_i^j) \quad (4)$$

Where w1, w2, and w3 are inverted indexing weight parameters, and their sum is 1.

A. Document semantic network generation for a document

Main motive behind the process of document semantic network generation for each document is to include as many semantically different word groups which leads to represent semantic meaning of document efficiently. A document semantic network, which contains a graph which connects all major terms from the document is constructed by an algorithm iteratively takes up a term one by one and builds the graph which is connected. As mentioned previously, brevity can

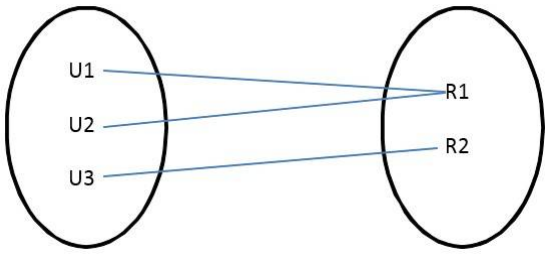


Fig. 2. Many-to-Many Mapping

omits some of the terms from a document, those terms must be considered in the process of measurement of relevance score for each document semantic network.User profile is built by considering majorly accessed document by user which can be used to contribute to personalized ranking results.As a result of this, an user profile indicates which information or properties user frequently considers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

V. SOFTWARE REQUIREMENTS SPECIFICATION

It specifies for the software requirement for the search the particular documents and various activities to be performed.

A. Software Requirements

- Windows 7 or later • java
- Eclipse or any IDE
- Spring Framework 4.0 and above
- Swagger Document
- Rule base json file • Implemented Api code files.

B. Hardware Requirements

- Intel(R) Core(TM) i5 CPU @ 2.90GHz or later, width : 64 bits
- Memory : 8 GB DDR3 or more
- Capacity : 1697MHz or more4
- Cores : 4 or more

VI. MATHEMATICAL MODEL

Process

Let us consider S as a system for CONCEPT BASED USER PROFILE.

S= {...} INPUT:

Identify the inputs

F= f1, f2, f3 .....,fn— F as set of functions to execute commands.

I= i1, i2, i3—I sets of inputs to the function set

O= o1, o2, o3.—O Set of outputs from the function sets

S= I, F, O

I = Data submitted by the user,

O = Output of desired data,

F = Functions implemented to get the output

φ = Failures and Success conditions.

Mathematical Model for proposed system

Proposed system S has following components:

S= {*start,end,X,Y,Ff,Fme,DD,NDD*}

Where, X = {*X*1,*X*2,*X*3}

Y = {*Y* 1,*Y*2}

Ff = {$F_{organize}$}

F*me*={F*preprocess,Fmajor*$_t$*erm,Fsdn,Frank*} $F_{preprocess}$= function which applies pre-processing methods on each input document

$F_{majorterm}$= function which selects major terms from all set of words of each document

Fsdn= function which forms graph i.e. semantic network of each document and query

F_r*ank* = function which calculates rank of document corresponding to each query.

DNN = {*X*}

NDD = {*Query*}

X is input set in which each item is :

*X*1 ={u e X1 — u is the set of input documents}

*X*2 ={u e X2 — u is the set of relationships considered} *X*3 ={u e X3 — u is the query fired by user } Y is output set in which each item is : *Y* 1 ={u e Y1 — u is the semantic network of each input page}

Y 2 ={u e Y2 — u is the semantic network of query}

Y 3 ={ *X*1,*X*3 e Y3 — u is the rank of each input document with respect to each query}

Failures:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

Huge database can lead to more time consumption to get the information.
Hardware failure.
Software failure.
Success:
Search the required information from available in Datasets. User gets result very fast according to their needs.

VII. RESULTS AND ANALYSIS

The main goal of the DSN generation process is to connect separated term-groups so as to represent the semantics of a document precisely. To make a DSN in which all major terms in the document are connected, the algorithm iteratively gathers a term one by one to make a connected graph. Since some of terms can be omitted in a document for brevity, those terms should be considered in the relevance measuring process. To provide personalized ranking results, we exploit users highly accessed documents to build a user profile. Using the algorithm, each of the DSNs for accessed documents are combined together to form a connected graph. Thus, the user profile, a portion of the ontology, represents which parts the user frequently handles and which properties the user mainly considers.
Following example will clear the idea of how the Document Semantic Network of text document will be generated:-

Example:- Tom is a cat.Tom caught a bird.Tom is owned by John.Tom is ginger in colour.Cats like cream. The cat sat on the mat.A cat is a mammal. A bird is an animal. All

Sr. No	Existing Method Result	Proposed Method Result
1	60%	80 to 85%

Fig. 3. Comparative Table

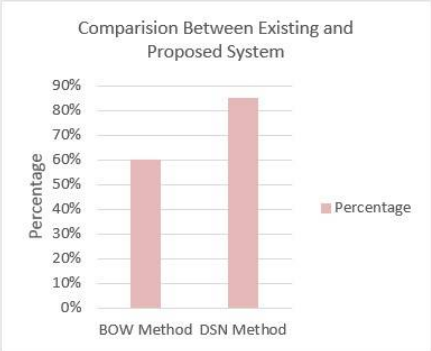


Fig. 4. Comparison of existing methods with proposed method

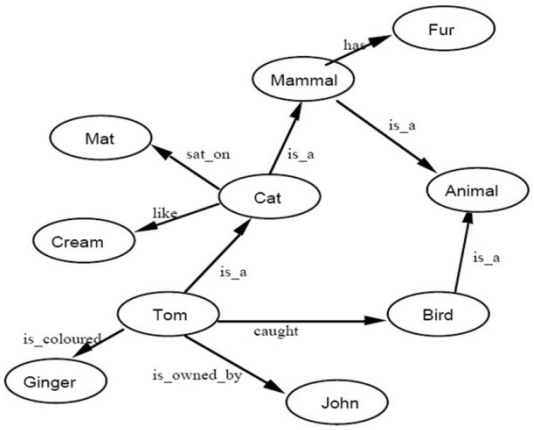
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018



VIII. COMPARATIVE RESULTS

In this section, comparisons between the proposed approach and other existing approaches are provided for performance evaluation. In Existing use bag-of-words method for document retrieval But in proposed work use Document semantic network for document retrieval.

IX. CONCLUSION

The basic idea behind this proposed work is to measure the impact or influence of relationships between terms in documents.Proposed method considers relationships between terms as an important factor for calculating the relevance score and ranking of the particular document.The proposed work provides mainly two advantages over the traditional approaches, which are: (1) Document semantic network enables to represent the core semantics of document which includes important terms and the relationship between them and (2) users' intent and interest are considered by exploiting user profile for calculating ranking function.Byconsidering these relationships more relevant document can be retrieved even if they they do not have exactly matching term with the user query.Users' satisfaction is improved by providing more accurate ranked and retrieval results relevant to query.

ACKNOWLEDGMENT

I take this opportunity to express my deep sense of gratitude for giving me this splendid opportunity to select and present this Dissertation topic.I would like to thank our Head Of Department Dr. R. B. Ingle for encouragement and providing me with the best facilities for my work. I thank all the sta members, for their indispensable support and for most valuable time lent as and when required.

REFERENCES

1] Nida Aslamb, Irfan Ullah , Jonathan Loo , RoohUllah , Martin Loomes , SemRank: ranking refinement strategy by using the semantic Intensity (2011)
[2] Chi Chen,Member, IEEE, XiaojieZhu,Student Member, IEEE, PeisongShen,Student Member, IEEE,JiankunHu,Member, IEEE, Song Guo,Senior Member, IEEE, ZahirTari,Senior Member, IEEE, andAlbert Y. Zomaya,Fellow, IEEE, An Efficient Privacy-Preserving Ranked Keyword Search Method(2016)
[3] Rajni Kumari Rajpal Mr.Yogesh Rathore, A Novel Techinque For Ranking of Documents Using Semantic Similarity(2014)
[4] Seema Rani, Upasana Garg2 ,1Guru Kashi University, Department of CSE, Talwandi Sabo, Punjab, India A, Ranking OfWeb Documents Using Semantic Similarity And Artificial Intelligence Based Search Engine(2014)
[5] Seema Rani, Upasana Garg2 ,1Guru Kashi University, Department of CSE,Talwandi Sabo, Punjab, India A, A Review Paper On Web Page Ranking Algorithms(2015)
[6] Pooja Arora, Prof. Om Vikas, Senior Member, IEEE India, Semantic Searching and Ranking of Documents using Hybrid Learning System and WordNet(2011)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 6, June 2018

[7] Ali I. El, DsoukyHesham ,A. AliRabab S. Rashed,Eygpt, Ranking Documents Based on the Semantic Relations Using Analytical Hierarchy Process(2016)
[8] F. Crestani, Exploiting the similarity of non-matching terms at retrieval time,Inf. Retrieval 2 (2000) 2747
[9] D. Metzler, W.B. Croft, A Markov random field model for term dependencies,in: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 05, ACM Press,New York, New York, USA, 2005, p. 472