ANALYSING AND IMPROVING QUALITY ASSURANCE IN DOCUMENT SEARCH-ENGINE INCORPORATING A DOCUMENT-RANKING ALGORITHM FOR TEXT-MINING

[1] Devendrasingh Thakore, [2]Dr. Akhilesh R. Upadhyay
[1]Research Scholar, JJTU, India
[2]Dept.of EC Engg., Sagar Institute of Research and Technology – Bhopal, 462041(M.P.), India
Email: deventhakur@yahoo.com, akhileshupadhyay@yahoo.com

Abstract
This paper discusses and revolves around providing an agent-based text-mining document search-engine architecture and implementing two algorithms one of which will be used for document-weighting and other which can be used for document-ranking. The main challenge lies in defining the factor on which the algorithms work and hence the 'weight of the document' has been defined as the deciding factor. While browsing the internet, people use different Search Engines such as Google, Yahoo, and Bing etc. In India Internet access is not easily available everywhere especially to the person who knows a little about computers. But there will be people who are especially seeking proper and genuine information related to their queries in a particular domain such as cancer diseases in medical science, data mining techniques in computer science, etc. Hence effort is made here is to develop a web-browser based document search-engine which incorporates novel algorithms for document-weighting and document-ranking and hence provide a good way to find out relevant documents with ease. With incorporating of Quality assurance activities in the software development phases ensures all required issues are addressed and implemented to achieve quality product development.

KEYWORDS: data-mining, text-mining, information retrieval, quality assurance at requirements analysis, quality assurance at design phase, quality assurance at implementation.

1. INTRODUCTION

The fundamental concepts involved in this paper are Data Mining, Text-mining, Intelligent agents, Document-weighting, Document-ranking Algorithms and Performance quality assurance. All these concepts and terms are the main base of this paper.

1.1 Data-mining

Data mining is the analysis of (often large) observational data sets to find unsuspected relations and to summarize the data in novel ways that are both understandable and useful to the data owner.

In other words data mining is a process of finding previously unknown, profitable and useful patterns hidden in data, with no prior hypothesis.

One more definition is "Data mining is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in data warehouses and by using pattern recognition technologies as well as statistical and mathematical techniques." Data mining provides a means of extracting previously unknown, actionable information from the growing base of accessible data in data warehouses to create competitive advantages for organizations.

It derives business intelligence from the data warehouse by using advanced analytical techniques such as neural networks heuristics, inductive reasoning, and fuzzy logic. Data mining applications are supported by a set of algorithmic approaches used to extract the relevant relationships in the data. The most commonly used approaches are: association, sequence-based analysis, clustering, classification, and estimation. The data mining process itself is organized into four major steps: data selection, data transformation, data mining, and result interpretation.

1.2 Text-mining

Text mining is a variation on a field called data mining. It sometimes alternately referred to as text data mining, roughly equivalent to text analytics which refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases or unstructured text. As the most natural form of storing information is *text*, text mining is believed to have a commercial potential higher than that of data mining. In fact, a recent study indicated that 80% of a company's information is contained in text documents. Text mining, however, is also a much more complex task as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining. [1]

Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the

algorithms and methods of KDD, data mining, machine learning and statistics. Thus, one selects a similar procedure as with the KDD process, whereby not data in general, but text documents are in focus of the analysis. From this, new questions for the used data mining methods arise. One problem is that we now have to deal with problems of — from the data modeling perspective— unstructured data sets. [2]

Text-mining is termed as the combination of various fields, as shown in Fig.1, namely Data Mining, Information retrieval (Extraction), Web Mining, Computational Linguistics & NLP, and Statistics.

Fig. 1 Conceptual study of Text-mining

DATA MINING: Data Mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information.

INFORMATION RETRIEVAL (EXTRACTION): Information extraction is the field encompassing the extraction of facts from structured texts such as databases or unstructured texts such as text documents.

WEB MINING: Web mining is the field which uses the data mining techniques to automatically discover and extract information from web documents and services such as online repositories. [4]

NATURAL LANGUAGE PROCESSING (NLP): The general goal of NLP is to achieve a better understanding of natural language by use of computers. Others include also the employment of simple and durable techniques for the fast processing of text, as they are presented. [2]

STATISTICS: Statistics is the study of collection, organization and interpretation of data. When this is used in accordance to text mining the data is unstructured text.

2. LITERATURE SURVEY

Intelligent agents play the role of assistants by allowing managers to delegate work that they could have done to the agent software. Agent technology is finding its way into many new systems, including decision support systems, where they perform many of the necessary decision support tasks formerly relegated to a uniquely human activity. Software agents are useful in automating repetitive tasks, finding and filtering information, intelligently summarizing complex data, and so on, but more importantly, just like their human counterparts, intelligent agents can have capability to learn from the managers and even make recommendations to them regarding a particular course of action. Agents possess several common characteristics, such as their ability to communicate, cooperate, and coordinate with other agents in a multiple agent system. Each agent is capable of acting autonomously, cooperatively, and collectively to achieve the collective goal of a system. The coordination capability helps manage problem solving so that co-operating agents work together as a coherent team. [1].Now, taking inspiration from this three different agents that can work collectively for 'Document-Search Engine' system. These three agents would be Key-order based agent; Keyword-sense based agent and an intelligent agent which incorporates the document-weighting and document-ranking algorithms. This research work is mainly concerned with the working of an intelligent agent.

Now, based on the above concept, the underlying architecture for the proposed 'Document Search Engine' system is developed. This research work is mainly concerned with the working of the Agent 3. The importance of good weighting methods in information retrieval is utmost vital, because the evidence is presented that good weighting methods are more important than feature selection process for the good performance in information retrieval.

There are few methods to calculate ad-hoc weights of the documents such as tf*idf method, and relevance feedback method. These methods are based on two important factors; one is statistical occurrence information and another is a history of how well this feature has performed in the past. In many situations, it is impossible to obtain history information and hence initial weights are often based purely on statistical information. [3]

TF*IDF METHOD

The tf*idf method (term frequency times inverse document frequency) assigns weight w_{ik} to term (word) T_k in document D_i, and in inverse proportion to the number of documents to which the term is assigned.

$$\text{Weight}(t) = tf * idf$$
$$= tf * \log(n/df)$$

Where, t = term to be weighted
tf = frequency of that term relative to other terms
n = total no. of documents
f = no. of documents in which the term 't' appears.

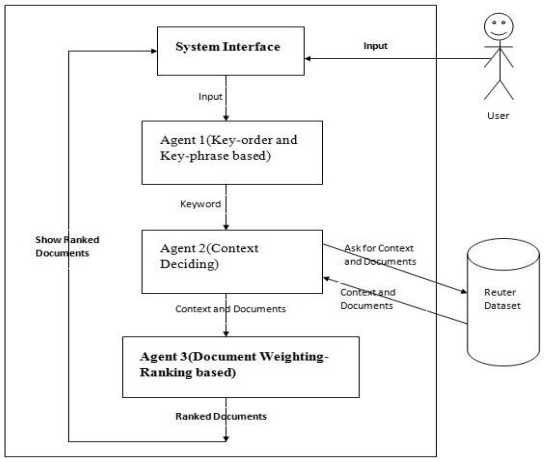

Fig.2 Underlying Architecture of the proposed systemT

TF*IDF METHOD used in Document-weighting *Algorithm*

The tf*idf method used in Document-weighting algorithm has little bit different significance than original one. Hence the equation is same, but the interpretation of terms used it little bit different.

$$\text{Weight}(d1) = tf(d1) * idf$$
$$= tf(d1) * \log(n/df)$$

Where, d1 = document to be weighted

tf = frequency(occurrences) of the particular term(inputted word) in documentd1.

n = total no. of documents

df = no. of documents in which the particular term (inputted word) appears.

Idea behind Document-Ranking Algorithm:

"After getting the weight of documents in which the inputted word appears, the average weight is calculated by doing sum of weights of all resulted documents and then by dividing them by the number of resulted documents. Then that average weight is taken as the pivot element and based on that pivot element two lists are created: one containing all the documents having higher weight than the average weight and another one having lower weight than the average weight. Then the list which contains the documents having higher weight is to be taken again and the average weight will be calculated for those documents. Thus the procedure needs to be repeated and so as in the end we will have the ranked documents."

3. METHODOLOGY

The work basically proposes an agent-based text-mining framework incorporating document-weighting and document-ranking algorithms. The quality assurance activities have been performed to assess the quality of work done. The working steps are as follows:

a) Users have to enter a word (constrained to the dataset, because they are news articles).

b) Based on inputted word 'Document-finding Algorithm' groups together the documents (articles) in which inputted word appears.

c) For those, documents which are grouped together in the previous step, the weight value have been calculated using 'Document-weighting Algorithm'.

d) Then the 'Document-ranking Algorithm' will be applied and the documents will be ranked and presented to the user.

Document-Finding Algorithm Steps:

a) Takes input word from user.

b) Initialize the counter variable to zero, which will indicate the no. of documents in which the inputted word appears.

c) Look for the documents, in the dataset, in which the inputted word appears.

d) Increment the counter by 1 when the word found in the document and add the document to the final list which will consist of final resulted documents.

e) Display the list to the user.

Document-Weighting Algorithm Steps:

a)Take the final list consist of final resulted documents.

b) Count the occurrences (frequency) of the inputted word in each and every document of the list.

c) Then apply the (tf*idf) equation to calculate weight of that document. Here (tf) represents term frequency and (idf) represents inverse document frequency.

d) Create one array consisting of weight of the resulted documents which are stored in final list.

e) Display the list of documents and their respective weight values to the user.

Document-Ranking Algorithm Steps:

a) Find the average weight value by adding the weights of all documents in 'final list' and dividing that by 'total documents in which word appears'.

b) For all the documents in 'final list', compare the weight of document with the 'avg. weight'.

c) The documents which are having less weight than the 'avg. weight' will be added to another list 'M' and their weights are added to another array 'weight1' and the documents which are having more weight than 'avg. weight' will be added to another list 'N' and their respective weights are added to another array 'weight2'.

d) If there are more than one documents in the list 'N' then recursively call the Algorithm with parameters related to documents in list 'N'. If all the documents in the list 'N' are having same 'weight value' then there is no need to repeat the process because every time the average weight will be same and thus it will go into the infinite loop. And if there is only one document in the list 'N' then add that document in the 'final' list.

e) If there are more than one documents in the list 'M' then recursively call the Algorithm with parameters related to documents in list 'M'. If all the documents in

the list 'N' are having same 'weight value' then there is no need to repeat the process because every time the average weight will be same and thus it will go into the infinite loop. And if there is only one document in the list 'N' then add that document in the 'final' list.
f) Display the documents in the list 'final' to the user.
Thus 'final' will have all the ranked documents with their title and data which will be displayed to user who had given the search query using particular keyword.

4. QUALITY ASSURANCE

Quality has been the important factor in each and every product or services that human-being uses. Quality plays an important role in convincing people to buy particular products or to use particular services. For ex., if one wants to buy a new mobile phone then apart from the features quality of the material used in manufacturing that mobile phone, quality of the accessories provided with the mobile phone, quality of the system software(s) installed into mobile etc. also plays an important role.
Quality assurance activities at three phases of SDLC:
i) Quality Assurance at Requirements Analysis Phase
ii) Quality Assurance at Design Phase
iii) Quality Assurance at Implementation Phase

4.1 Quality Assurance at Requirements Analysis Phase

Quality assurance is mainly performed onto two kinds of stated requirements specifications:
A) Functional requirements specification
B) Non-Functional requirements specification
A) Functional requirements specification
Functional requirements specification is an important document in the SD life cycle because it provides the insight to the overall functionality related issues of the system. So, performing quality assurance related tasks at this stage provide us the more controlled, managed and refined way to achieving good software quality. We can simply review each and every requirement and its different aspects onto software development and then we can prepare quality assurance checklist document, containing questionnaire regarding functional requirements at different levels.

a) Pre-requisites
Pre-requisites are the important resources and tools required before the actual development starts. Pre-requisites for this project work are:
Reuters-21578 News articles dataset.
JDK 1.5 (or above) is required at the development stage to run the JAVA programs.
Apache web server is required to run the developed JSP/Servlet APIs.
Any web-browser is required to run the developed JSP/Servlet APIs.

b) Input level requirements
These are the requirements regarding inputs to the system developed.
Reuter 21578 SGML/XML News Articles Dataset.
Keywords/words from the Dataset files.

c) Output level requirements
These are the requirements regarding outputs from the system developed.
Documents (News articles) containing the 'searched keyword'.
Weights of the documents found, containing 'searched keyword'.
Ranked documents according to weights
All these should be in a table-format display.

d) Procedural (Workflow) requirements
These are the procedural or workflow steps for the system developed. The system should take the input (in form of keyword) The system should find out the documents (news articles) in which inputted word/keyword appears. The system must use the Reuters-21578 news articles dataset as the resource to find out the documents. The system should perform the task of 'weighting documents' for the resulted documents. The system should perform the task of 'ranking documents' for the resulted documents.
The ranking is performed based on weights of the documents. Ranked documents are presented to the user in form of table-form display.
B) Non-Functional requirements specification
Non-functional requirements specification is a handful document, in accordance with functional requirements specification, as it addresses the issues regarding non-functional but vital requirements. Here we are mainly dealing with the constraints at different levels of the system development. We can simply review each and every non-functional requirement and its different aspects onto software development and then we can prepare quality assurance checklist document, containing questionnaire regarding functional requirements at different levels.

a) Dataset Constraints
Currently the system works on the Reuater-21578 news articles dataset, but in future we can take some other datasets or may we can take this system into Internet environment. Currently the dataset used in the system as a resource is in SGML/XML form. In future, other formats may also be supported.
b) Time Constraints
The whole work needs to be completed within stipulated time period. Proper plan and time-adhered schedule had been prepared to complete the tasks in time.

This document is used as baseline for inspection. Quality assurance activities are performed while working on above phase by reviewing each requirement with strictly checkmarks. System functionalities are checked and added addressed suggestions to minimize defect or clarification against the Functional Specification document. Checklists document for the web-application is prepared which

provides a complete list of items to be verified and also provide space for documenting findings of the checks performed.

Sr. No	Topic/Requirement	Yes	No
I	**The Document**		
1	Is the document prepared according standard template?	*	
2	Are there any ambiguities in the prepared document?		*
3	Is prepared document complete?	*	
4	Does the document address the issues regarding functional and non-functional requirements of the system?	*	
II	**Functional Specification**		
1	Are all the required functionalities are properly defined?	*	
2	Are the functional requirements defined are clear and unambiguous?	*	
3	Are the inputs to the system conforming to functional requirements?	*	
4	Are GUI interfaces are fully defined and addressed properly?	*	
5	Is the flow of system conforming to the stated functional requirements?	*	
6	Are desired time complexity and space complexity calculations have been performed for the algorithms developed?	*	
7	Is the 'weighting document' task is properly performed?	*	
8	Is the 'ranking document' task is properly performed?	*	
9	Is the developed system supports any other dataset currently?		*
III	**Non-Functional Specification**		
1	Does the 'Time line chart' or 'Time-schedule' plan have been prepared for the system development?	*	
2	Does the system development task is feasible to complete within defined time limit?	*	
3	Does the system development conform to dataset constraints?	*	
4	Does the system require assistance from any other resources rather than those define earlier?		*

Table 1 : Checklist Document at Requirements Analysis Phase

From the above table, containing checklist document at requirement analysis phase, we can state following things:

1) The checklist document addresses various issues related to functional and non-functional requirements of the system using questionnaire form template.

2) As well as it also addresses the issues regarding achieving good quality at requirement analysis level using the form of questionnaire prepared to conform to quality measures.

3) There are few 'NOs' and mostly 'YESs', so the checklist document provides a positive outcome to the good quality achievement plan.

4.2 Quality Assurance at Design Phase

The design phase of the SDLC is a very important phase because it details the functionality oriented design and provides the significant insight onto how actually the things will work in the implementation phase. The design phase provides the layouts of the user interfaces of the system and also provides the insights onto functional components of the system at design level. The design engineers normally need to run the check on traceability of design with requirements specified.

The primary quality assurance activity during the design phase could be the formal review of the preliminary and detailed design documents. These documents are verified for their consistency, completeness, and correctness within themselves and with the requirements specification document.

Design review by checklist document

The checklist document contains the type of questionnaire regarding different issues focusing on different components of the design phase. The checklist document for design phase of the thesis work, titled, 'Text-Mining Based Document Search-Engine Incorporating a Document-Ranking Algorithm', is shown as below:

Sr. No	Topic/Requirement	Yes	No
I	**Requirement Traceability**		
1	Does the design adhere to the requirements specified?	*	
2	Does the design have any ambiguities respective to requirements?		*
3	Does the design artifacts traceable to requirements?	*	
II	**Functionality**		
1	Do the design artifacts adhere to functionality issues?	*	
2	Are all the design artifacts are clear	*	

	and concise according to functionality demanded?		
3	Do design artifacts provide a new kind of architecture for the system?	*	
4	Does the design support any other operating environment than suggested?		*
5	Are all the user interfaces, data flows and control flows are properly addressed?	*	
III	**Design Centered**		
1	Do the design artifacts provide a new design arena that leads the system to easy and concise user interface?	*	
2	Does the design provide clear and concise understanding of how the system could be workable?	*	
3	Does the design support any other operating architecture than suggested?		*
4	Does the design follow standard techniques to describe system?	*	
IV	**Design Feasibility and Reliability**		
1	Does the design adhere to time, cost and effort feasibility constraints of the system?	*	
2	Do the time and space complexity of the designed algorithms related issues have been addressed?	*	
3	Does the impact of undesired events, such as system restart or system failure, have been considered?		*

Table 2 : Checklist Document at Design Phase

The checklist document prepared above shows that all the required issues related to design are considered and addressed properly resulting into the good quality at design phase.

4.3 Quality Assurance at Implementation Phase

Implementation phase of SDLC is the phase where the actual coding has been performed or to be specific the actual software is developed through coding according to requirements collected and designs created during earlier SDLC phases. Before the actual implementation or coding starts all the requirements are collected, reviewed and documented properly and same as all the design artifacts have been created according to stated requirements and documented properly. So, implementation can be considered as the more on like mechanical task rather than the logical task.

Different types of Code-Reviews, Code-Inspections, Active Design Reviews and Walkthroughs have been found to be very useful in improving the quality of software. Code inspections should check for technical accuracy and completeness of the code, and to verify that it implements the planned design, and ensure good coding practices and standards are used.

Implementation review by checklist document

The checklist document contains the type of questionnaire regarding different issues focusing on different components of the implementation phase. The checklist document for implementation phase of my thesis work, titled, 'Text-Mining Based Document Search-Engine Incorporating a Document-Ranking Algorithm', is shown as below:

Sr. No.	Topic/Requirement	Yes	No
I	**Interfaces ,Code and data**		
1	Does the implementation adhere to design artifacts?	*	
2	Does the code (programs) are traceable to design artifacts suggested?	*	
3	Does the coding have been performed according to standards, such as placing comments so that anyone can understand the code properly?	*	
4	Are there any ambiguities in code regarding the design artifacts?		*
5	Does the code include all the required libraries or packages?	*	
6	Do the function calls and returns work properly in the code?	*	
7	Are requests to allocate memory checked for success before attempting to write to that memory?		*
8	Are there any unidentified jumps or loops present in the code?		*
9	Are the required variables declared properly?	*	
10	Are the required variables initialized at the beginning of the main () function or other functions?	*	
11	Does the system perform required functions properly?	*	
12	Does the system provide required output in the required format?	*	
II	**Feasibility and Reliability**		
1	Does the design adhere to time, cost and effort feasibility constraints of the system?	*	
2	Does the backup have been maintained for developed code?	*	
3	Does the output provided by the system is reliable?	*	
4	Is the code developed is portable onto other computer systems?	*	
5	Does the impact of undesired events, such as code malfunctioning or system failure, have been considered?		*

Table 3 :Checklist Document at Implementation Phase

The checklist document prepared above shows that all the required issues related to implementation are considered and addressed properly resulting into the good quality at implementation phase.

5. CONCLUSION

The ranking of documents is very important part of the document-search engine systems as it provides the easier way to find out relevant documents to the users' query. The system uses agents-based architecture and incorporates algorithms for document-weighting and document-ranking. The incorporating of quality assurance activities at the three phases namely: Requirement Analysis, Design Phase and Implementation Phase ensure that all the required issues relating to software development are addressed and implemented. This helps in assessing the quality of work done and ultimately results in achieving a quality product in term of development and performance. The system (web browser based application) developed will provide an easier way for document search and those documents are presented in the ranked form.

6. REFERENCES

1. "IDM: An Intelligent Software Agent Based Data Mining Environment" by Ranjit Bose and Vijayan Sugumaran.
2. "A Framework of an Automated Data Mining System Using Autonomous Intelligent Agents" by J.Rajan and Dr. V.Sarvanan
3. "Importance of proper weighting methods" by Chris Buckley
4. "The Anatomy of a Large-Scale Hypertextual Web-Search Engine" by Sergey Brin and Lawrence Page.

AUTHORS BIOGRAPHY

Devendrasingh Thakore is pursuing Ph.D form JJTU, India. He obtained M.E. (Comp.) degree from the BU, Pune in 2004, M.B.A. (Marketing.) from MITSOM, Pune and B.E. (Comp Engg) from Walchand College of Engg, Sangli [M.S.] in the year 1996 and 1990 respectively. His areas of interest are Computer Network, Software Engineering and Database System. He has seventeen years experience in teaching and research. He has published more than twenty research papers in journals and conferences. He has also guided ten postgraduate students.

DR. AKHILESH R. UPADHYAY obtained Ph.D. degree from the Swami Ramanand Teerth Marathwada University, Nanded in 2009, M.E. (Hons.) and B.E. (Hons.) in Electronics Engineering from S.G.G.S. Institute of Engineering & Technology, Nanded [M.S.] in year 2004 and 1996 respectively. He is currently working as Vice Principal and Head of Electronics and Communication Engineering Department at Sagar Institute of Research and Technology, Bhopal, India.

He has more than 12 years teaching and 3 years of industry experience. He is Associate Editor of Journal of Engineering, Management & Pharmaceutical Sciences, Ex-Editor of International Journal of Computing Science and Communication Technologies and member of editorial boards/review committee of various reputed journals and International conferences. He has more than 50 research publications in various international/national journals and conferences; he also authored more than 16 text/reference books on electronics devices, instrumentation and power electronics. He is recognized Ph.D. Supervisor for various Universities in India and presently guiding 11 Ph.D. scholars.