DESIGN OF TEXTUAL PRESENTATION FROM ONLINE INFORMATION USING HYBRID APPROACHES

S. Saraswathi
Department of Information Technology, Pondicherry Engineering College, Pondicherry, India
E-mail: swathi@pec.edu

Abstract
This paper describes the design of online textual presentation tool developed by using various approaches like information retrieval, information extraction, topic identification and sentence weightage algorithms. To design a presentation for the given user query, the following approaches are used: Information extraction based on Boolean and probabilistic approaches, Topic identification based on semantic based approach, Sentence weightage algorithm based on importance of the terms and topics involved in the documents. A combination of these approaches led to the design of PowerPoint presentation file.

Keywords:
Recall, Precision, Textual presentation, Topic Identification, Sentence Weightage, Summarization

1. INTRODUCTION

Information retrieval is the science and art of locating and obtaining documents based on information needs expressed to a system in a query language. The Boolean retrieval is the most simple of these retrieval methods and relies on the use of Boolean operators. The terms are linked together using AND, OR and NOT operators to the user query. A comparison of the Boolean and probabilistic information retrieval systems was done by Robert M.Losee [1]. The approach of comparing Boolean and term weightage models has become possible because of the development of analytic models of information retrieval performances. A probabilistic system ranks the documents for retrieval by assigning a numeric value to each document, based on the weights for query terms and the frequencies of the terms occurring in the documents. The models of probabilistic retrieval [2] provides researchers with a decision rule stating that a document should be retrieved if a calculated value that is based on several parameters is less than the cost based value. A text partition model [3] was proposed to determine the boundaries of discourse structures. It was based on association of noun-noun relations and noun-verb relations defined on discourse level and sentence level respectively. Three factors are considered for topic identification:1) repetition of words, 2) importance of words and 3) collocation semantics. A window was moved from the first sentence to the last sentence and the association norms for sentences in the current window were calculated. Hierarchically organizing data [4] according to topics has been accepted as a very successful method for organizing or browsing a large volume of textual documents in information systems. Such a topic hierarchy is used as a way of systematically organizing a large document collection; incoming documents (ie. Extracted from web) are indexed on topic hierarchies, which are presented as browsable directories to users with information needs. Construction of topic hierarchies is of great importance in information systems to organize huge numbers of online text documents. Youngjoong Ko [5] has described a method to improve the text categorization using the importance of sentences. The features from important sentences should be given weightage. The authors used two kinds of text summarization techniques: one uses the title and the other uses the importance of terms. This paper discusses on the design of a textual presentation application from online information using the various techniques related to information retrieval, topic identification, and sentence weightage.

This paper is structured as follows: section 2 and section 3 provides the overall design and module description of the online presentation, section 4 and 5 discusses on results and future enhancements.

2. DESIGN OF ONLINE PRESENTATION SYSTEM

Different techniques exist for information retrieval, topic identification and sentence weightage methods [6]-[9]. In the proposed system the combination of these approaches are used for designing a textual presentation tool. Fig.1 shows the overall design of online presentation and the steps needed for the design are as follows:

The user query will be send to the search engine. Documents related to the query in various file formats are collected and converted to text files. The contents of text files are checked for relevant information using information extraction techniques based on Boolean and probabilistic approaches. The topics are identified for the relevant text files using semantic based approach. Based on the sentence weightage, the presentation file is generated to the user based on the number of slides he/she requires in the Software Engineering domain.

3. MODULE DESCRIPTION

3.1 COLLECTION OF RELEVANT DOCUMENTS

The block diagram for collecting the documents is shown in the Fig.2. The user sends the required query to the search engine. The search engine will retrieve the information related to the query through the WebConversation and WebResponse classes in HttpUnit package. HttpUnit is a set of classes for automating browser functions. HttpUnit is a free, open source Java API for accessing web sites without a browser. The three main classes in the HttpUnit packages are WebConversation, WebResponse and WebRequest. The center of HttpUnit is the WebConversation class, which takes the place of a browser talking to a single site. The user has to send the needed query through WebRequest and file type to download. WebRequest is the abstract base class for the .NET framework's request or

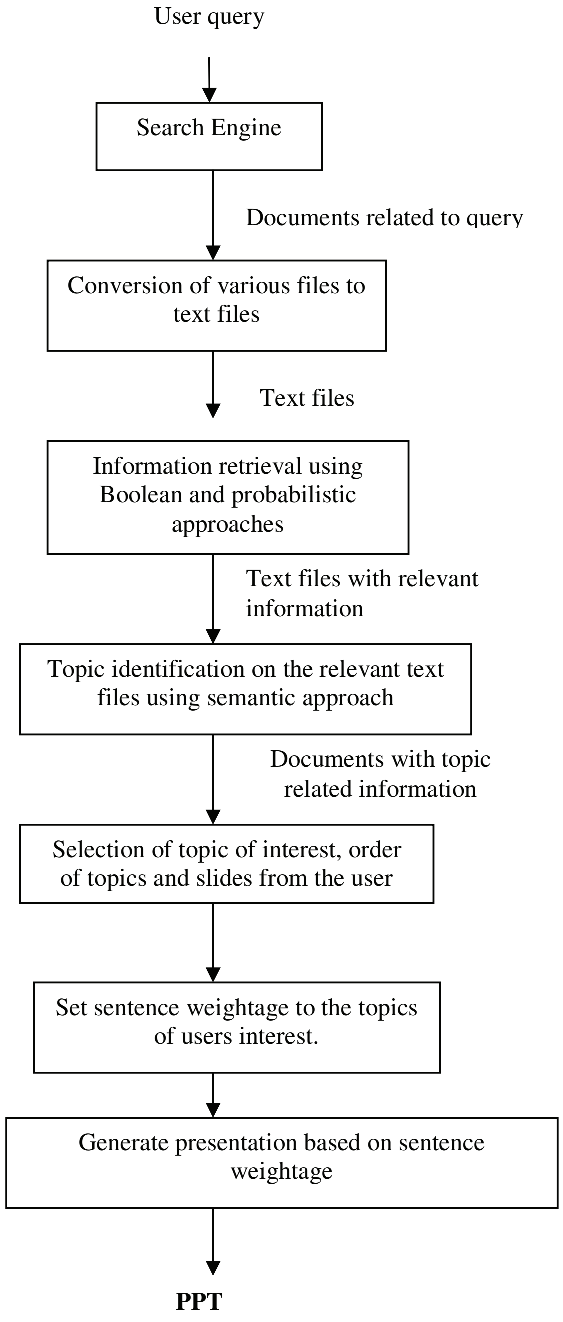

User query
Search Engine
Documents related to query
Conversion of various files to text files
Text files
Information retrieval using Boolean and probabilistic approaches
Text files with relevant information
Topic identification on the relevant text files using semantic approach
Documents with topic related information
Selection of topic of interest, order of topics and slides from the user
Set sentence weightage to the topics of users interest.
Generate presentation based on sentence weightage
PPT

Fig.1. Block diagram of online presentation

response model for accessing data from the internet. Web response is defined as a response to a web request from a web server. In the requested page the needed information and the file type are searched and retrieved. The retrieved files in various formats are stored in the user specified folder.

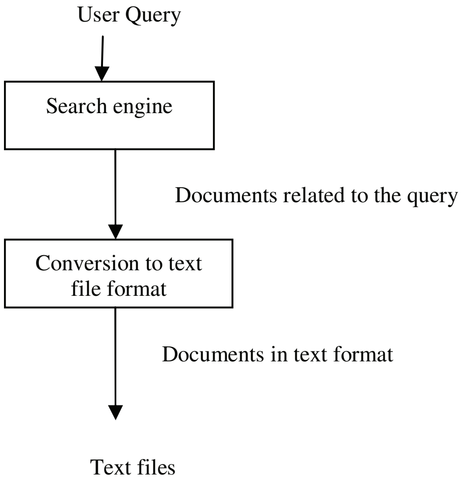

User Query
Search engine
Documents related to the query
Conversion to text file format
Documents in text format
Text files

Fig.2. Block diagram for collecting documents

In order to avoid the complexity of extracting information from various file formats, all the files should be converted to text file format. To convert various file formats to text file format, the tool named convert doc is used. Convert doc tool converts file formats like DOC, HTML, RTF, PDF, to TXT format. By increasing the search page number in google, the number of downloaded documents can be increased. Converted text files are downloaded in the user specified folder.

3.2 ELIMINATING IRRELEVANT INFORMATION

The block diagram for eliminating the irrelevant information is shown in the Fig.3. The converted text files might contain irrelevant information like references, acknowledgement, etc.

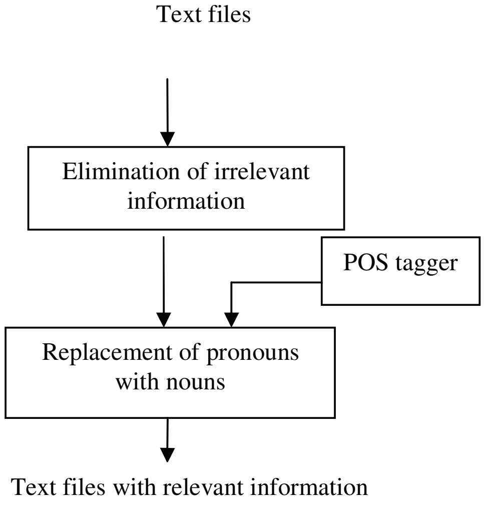

Text files
Elimination of irrelevant information
POS tagger
Replacement of pronouns with nouns
Text files with relevant information

Fig.3. Block diagram for eliminating irrelevant information

The irrelevant information is removed by detecting the keywords like acknowledgement and references. The garbage data which follows under the diagram are to be removed by detecting the keyword figure.

The authors with their affiliations part are removed by detecting the keyword references. The regular expression in java is used to compare the keywords in this module. A regular expression is a pattern of characters that describes a set of strings. Regular expression matching also allows to test whether a string fits into a specific syntactic form, such as an email address. The information which follows under the keyword references is deleted. The text files with relevant information are tagged using the Wintree POS tagger [10]. In the tagged file the pronouns of a sentence are replaced by the corresponding noun. The resulting text file of this module contains only relevant information.

3.3 TOPIC ASSIGNMENT

The block diagram for topic assignment is shown in the Fig.4. To assign the topic for each paragraph the maximum occurrence of keywords (nouns and verbs) and their semantically equivalent words is determined. The maximum occurred keywords are determined. The topics related to Software Engineering are stored in a database. The maximum occurred keywords in the paragraphs are compared with the topics stored in database to assign the suitable topics for them. Paragraphs with similar topics available in different text file are grouped together.

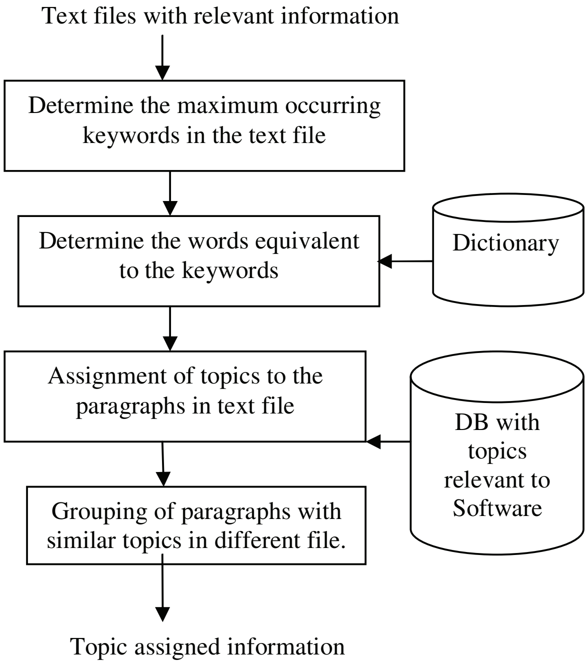
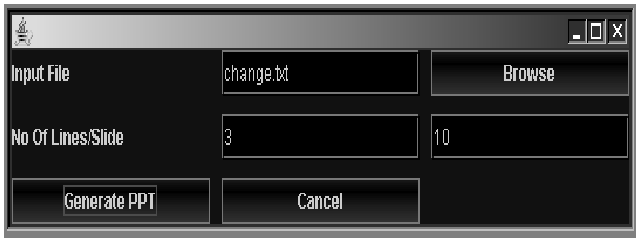

Text files with relevant information
Determine the maximum occurring keywords in the text file
Determine the words equivalent to the keywords
Dictionary
Assignment of topics to the paragraphs in text file
DB with topics relevant to Software
Grouping of paragraphs with similar topics in different file.
Topic assigned information

Fig.4. Block diagram for topic assignment

The next step deals with removal of semantically related sentences, within each topic assigned information. The percentage of similarity among sentences is calculated based on Euclidean distance measure of the keywords and the semantically related words present in them. If the comparison percentage exceeds the threshold value (70%), then the sentences are considered to be similar. The matched sentence will be removed. The size of the file gets reduced after the removal of the semantically related sentences.

3.4 GENERATION OF PPT

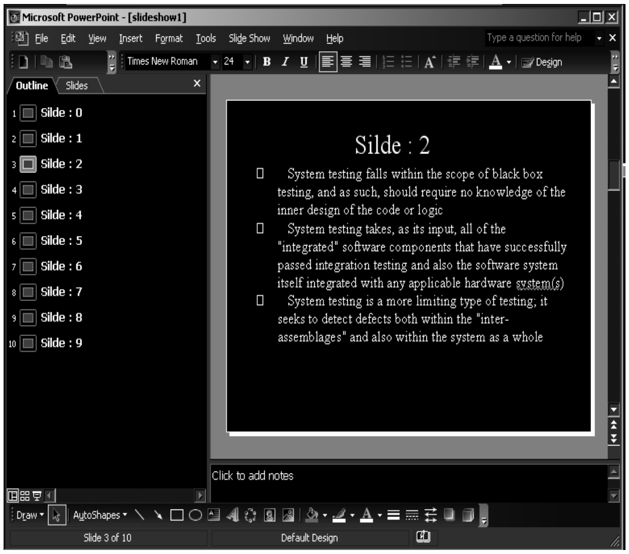
The topics and the sentences related to each topic are listed to the user with the line number. The user can list the topic order for generating the presentation. He can also list the order for arranging the sentences within each topic. Based on the user's interest the topics and their contents are re-ordered. The screenshot for user interface for PPT generation is shown in Fig.5. The number of the slides and the number of the sentences per slide is given by the user. The generated PPT file is shown in Fig.6.



Input File change.txt

No Of Lines/Slide 3 10

Fig.5. Screenshot for user interface for PPT generation



Silde : 2

- System testing falls within the scope of black box testing, and as such, should require no knowledge of the inner design of the code or logic
- System testing takes, as its input, all of the "integrated" software components that have successfully passed integration testing and also the software system itself integrated with any applicable hardware system(s)
- System testing is a more limiting type of testing; it seeks to detect defects both within the "inter-assemblages" and also within the system as a whole

Fig.6. Generation of PPT

4. RESULTS AND DISCUSSION

Precision and Recall are two widely used measures for evaluating the quality of results. Precision is defined as the fraction of the retrieved sentences that are relevant, and Recall is defined as the fraction of relevant sentences that are retrieved. The hash table is used to store the most relevant and irrelevant sentence based on the users interest in preparing the presentation. The precision and recall is calculated by using the equations (1) and (2).

$$Precision = \frac{No.of\ relevant\ sentence\ retreived}{No.of\ sentence\ retreived} \qquad (1)$$

$$Recall = \frac{\begin{array}{c} No.of\ relevant\ sentence\ retreived \\ based\ on\ users\ interest \end{array}}{No.of\ relevant\ sentence\ retreived} \qquad (2)$$

The result is analyzed for the Software Engineering domain. Software Engineering consists of the following topics namely Introduction to Software Engineering, Process models, System engineering, Design engineering, Project scheduling, Software testing, Requirements engineering, Software metrics, Web engineering, Component based software development, etc.. The topics that are chosen for the analysis of result are Software testing, Software metrics and Component based software development (CBSD).

The number of documents that are collected for Software Engineering was 100, out of which 20 documents were relevant to the software testing topic. The average size of those 20 documents is 53KB. The number of topics assigned for the 20 relevant documents is 40. The following are the some of the topics assigned in Software testing: System testing, Software testing, Integration testing, Black-box testing, White-box testing, Fault based testing, etc. The topics that are chosen for analysis are Unit testing, Acceptance testing, Performance testing, System testing.

Table.1. Precision and recall values for three domains

Domain Name	No.of relevant doc retrieved	Name of the Topic	No. of Sentences relevant to the topic	User 1		User 2	
				Precision	**Recall**	**Precision**	**Recall**
Software testing	20	Unit testing	40	0.75	0.91	0.8	0.75
		Acceptance testing	42	0.75	0.8	0.7	0.75
		Performance testing	44	0.92	0.54	0.6	0.8
		System testing	45	0.8	0.68	0.7	0.75
Software metrics	20	Software quality metrics	40	0.9	0.61	0.85	0.7
		Defect density metrics	40	0.8	0.43	0.75	0.8
		Customer problem metric	42	0.8	0.78	0.8	0.68
		Object oriented metrics	44	0.91	0.6	0.6	0.8
Component-based software development	20	Domain engineering	42	0.8	1.0	0.92	0.8
		Component reuse	40	0.8	1.0	1.0	0.77
		Overview of component based software development	44	0.8	0.65	0.65	0.8

The number of documents that were relevant to the software metrics topic was 20. The average size of those 20 documents is 51KB. The number of topics assigned for the 20 relevant documents is 37. The following are the some of the topics assigned in Software metrics: Software quality metrics, Product quality metrics, Function-based metrics, Operation-oriented metrics, Customer problem metrics, etc. The topics that are chosen for analysis are Software quality metrics, Defect density metrics, Customer problem metrics, Object oriented metrics.

The number of documents that were relevant to the CBSD is 20. The average size of those 20 documents is 48KB. The number of topics assigned for the 20 relevant documents is 23. The following are the some of the topics assigned in CBSD: Overview of CBSD, Domain engineering, Component adaptation, Component qualification, Component engineering, etc. The topics that are chosen for analysis are Overview of CBSD, Domain engineering, Component reuse.

Queries related to the above topics were given as input by the users. The sentences related topics were grouped together based on several topics. The order for presenting the presentation is provided by the user. The quality of the documents is measured by using precision and recall as shown in Table 1.

The result is tested with 50 users. The users can generate the presentation based on their order of interest. Thus the variation in the precision and recall values may occur by generating the presentation on their interests. The precision and recall values are plotted for the 80% of the user as shown in the Fig.7, 8 and 9 for the various topics in Software Engineering. The remaining

ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2010.0017

20% of the users got the precision and recall values as shown in the Fig.10, 11 and 12. The graphs are drawn by taking three domains like software testing, software metrics and component based development. They show variation because of the changes in various users' interests to generate presentation. The general inferences of the above graphs for the three domains are as follows:

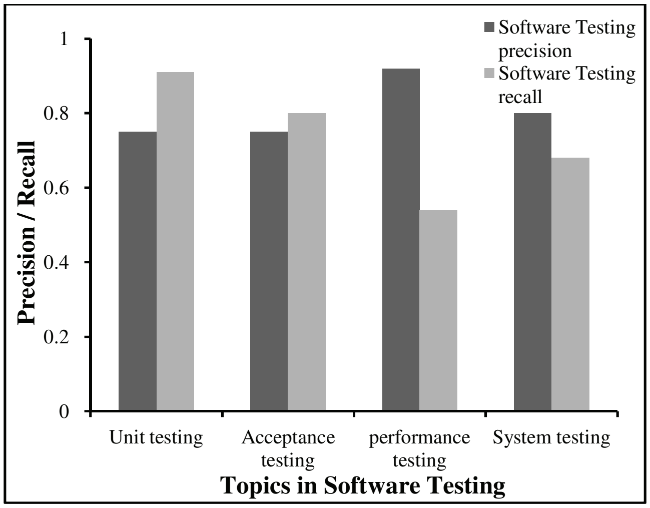


Fig.7. Precision & recall graph for 80% of users (user1) in the topic Software Testing

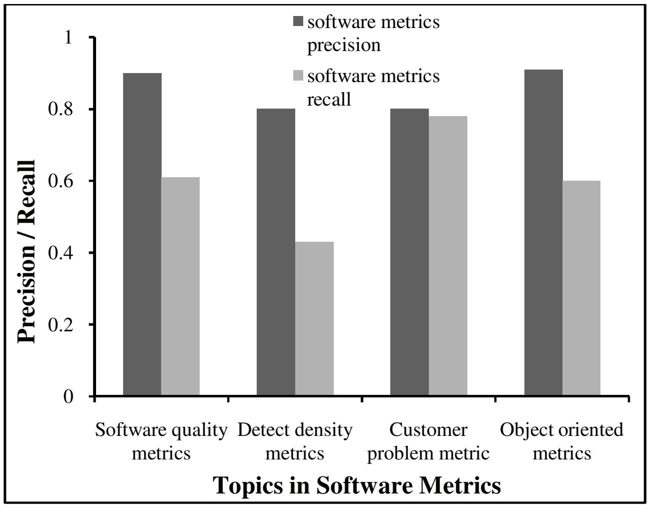


Fig.8 Precision & recall graph for 80% of users (user1) in the topic Software Metrics

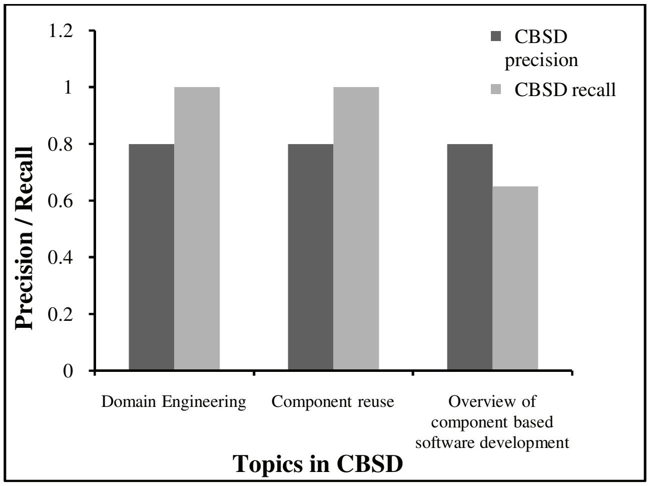


Fig.9. Precision & recall graph for 80% of users (user1) in the topic Component Based System Design

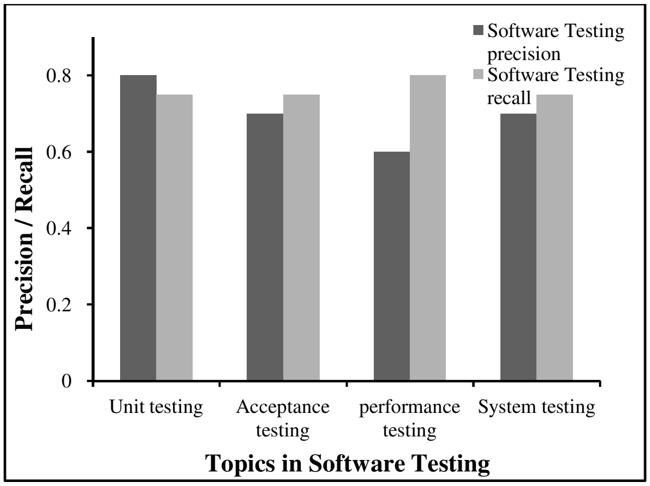


Fig.10. Precision & recall graph for 20% of users (user2) in the topic Software Testing

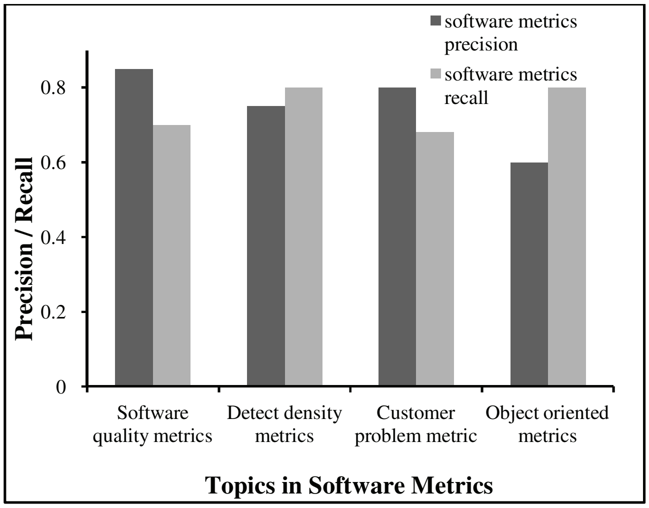


Fig.11. Precision & recall graph for 20% of users (user2) in the topic Software Metrics

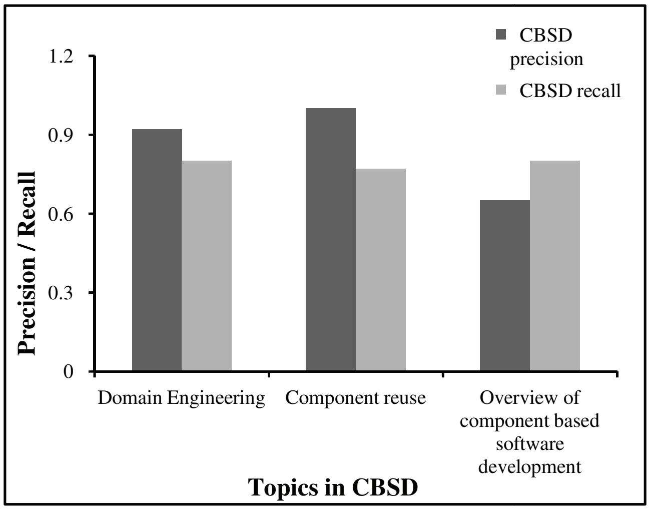
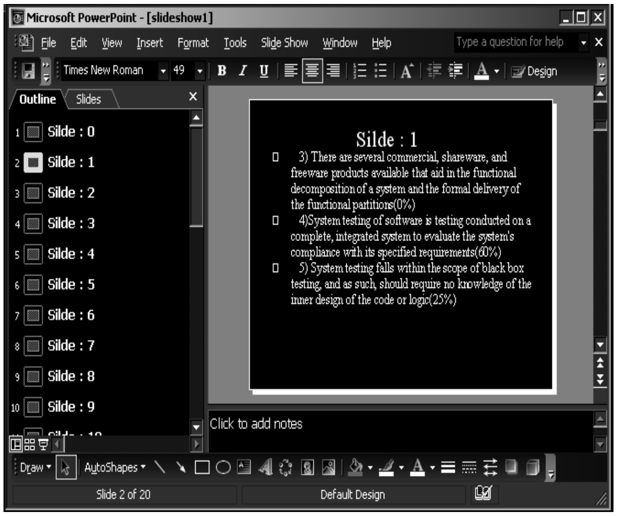


Fig.12. Precision & recall graph for 20% of users (user2) in the topic Component Based System Design

4.1 SOFTWARE TESTING

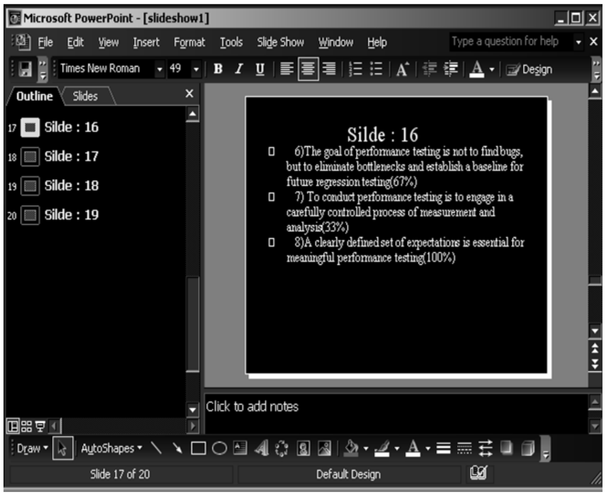
The four topics plotted in this domain are unit testing, acceptance testing, performance testing and system testing. The precision and recall value for unit testing to about 80% of the users varies by less difference, and the recall value is greater than the precision, as the topic contains more number of most relevant sentences. Acceptance testing also shows the same variation for user1. Performance testing shows a large variation with respect to precision and recall for user1. The precision value is greater than a recall since it contains relevant sentences count greater than the most relevant sentences. The system testing showed less variation in between unit testing and performance testing topics for user1. The precision value is greater than a recall since it contains relevant sentences count greater than the most relevant sentences. The relevancy can be increased by adding more relevant documents. The average precision and recall for user1 for the topic software testing is 0.8 and 0.73. The average precision and recall for user2 for the topic software testing is 0.7 and 0.76. The PPT generated by the users in their order of interests for the software testing is shown in the Fig.13 and 14.



Silde : 1

- 3) There are several commercial, shareware, and freeware products available that aid in the functional decomposition of a system and the formal delivery of the functional partitions(0%)
- 4)System testing of software is testing conducted on a complete, integrated system to evaluate the system's compliance with its specified requirements(60%)
- 5) System testing falls within the scope of black box testing, and as such, should require no knowledge of the inner design of the code or logic(25%)

Fig.13. PPT generated by user1 for the Software Testing topic



Silde : 16

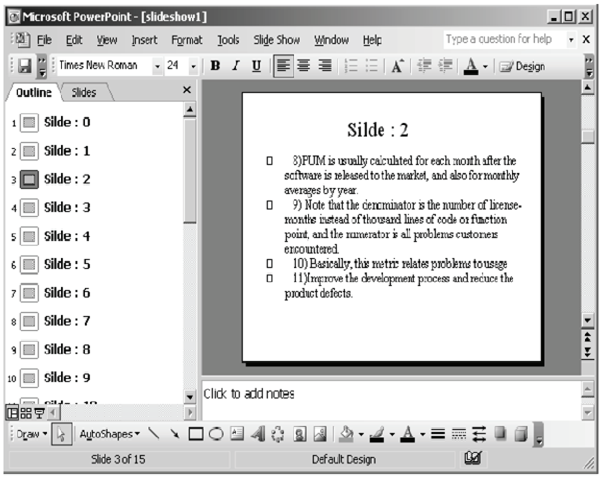
- 6)The goal of performance testing is not to find bugs, but to eliminate bottlenecks and establish a baseline for future regression testing(67%)
- 7) To conduct performance testing is to engage in a carefully controlled process of measurement and analysis(33%)
- 8)A clearly defined set of expectations is essential for meaningful performance testing(100%)

Fig.14. PPT generated by user2 for the Software Testing topic

4.2 SOFTWARE METRICS

The four topics plotted in this domain are software quality metrics, defect density metrics, customer problem metrics, and object oriented metrics. The software quality metrics shows a large variation with respect to precision and recall for user1. The precision value is greater than a recall since it contains relevant sentences count greater than the most relevant sentences. The defect density metrics also shows the same variation as software quality metrics for user1. The Customer problem metrics shows less variation for user1 since it contains slightly more number of relevant sentences. Object oriented metrics shows the precision value is greater than the recall values for user1 because of the more number of relevant sentences. The average precision and recall for user1 for the topic software metrics is 0.86 and 0.61. The average precision and recall for user2 for the topic software metrics is 0.75 and 0.75. The PPT generated by the users in their order of interests for the software metrics are shown in the Fig.15 and 16.

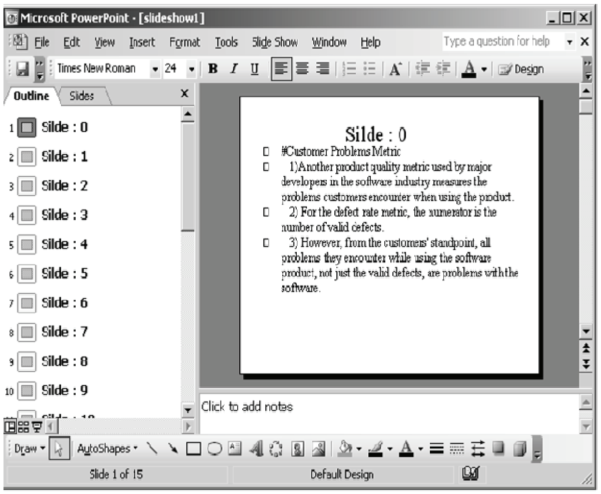
ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2010.0017



Silde : 2

□ 8)PUM is usually calculated for each month after the software is released to the market, and also for monthly averages by year.
□ 9) Note that the denominator is the number of license-months instead of thousand lines of code or function point, and the numerator is all problems customers encountered.
□ 10) Basically, this metric relates problems to usage
□ 11)Improve the development process and reduce the product defects.

Fig.15. PPT generated by user1 for the Software Metrics topic



Silde : 0

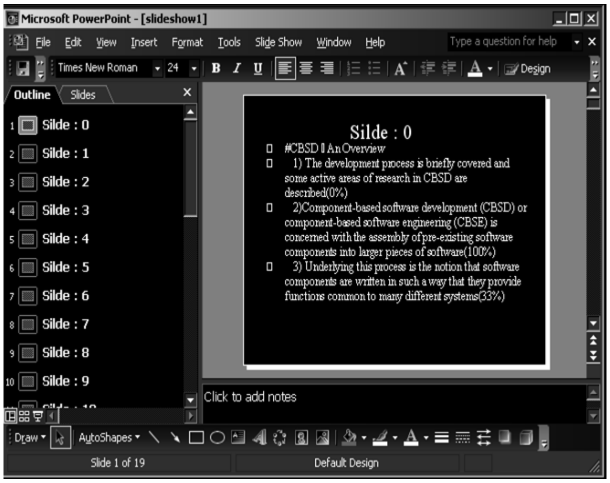
□ #Customer Problems Metric
□ 1)Another product quality metric used by major developers in the software industry measures the problems customers encounter when using the product.
□ 2) For the defect rate metric, the numerator is the number of valid defects.
□ 3) However, from the customers' standpoint, all problems they encounter while using the software product, not just the valid defects, are problems with the software.

Fig.16. PPT generated by user2 for the Software Metrics topic

4.3 COMPONENT BASED SOFTWARE DEVELOPMENT (CBSD)

The topics plotted in this domain are domain engineering, component reuse and overview of CBSD. The first two domains contain same precision and recall values for user1. The recall value for the first two domains is 1 and it shows that most of the sentences are most relevant and only very few sentences are irrelevant. The first two topics contain no relevant sentences. Whereas for the third topic the relevant sentences and irrelevant sentences count are equal to the most relevant sentences count and hence it contains less recall value.

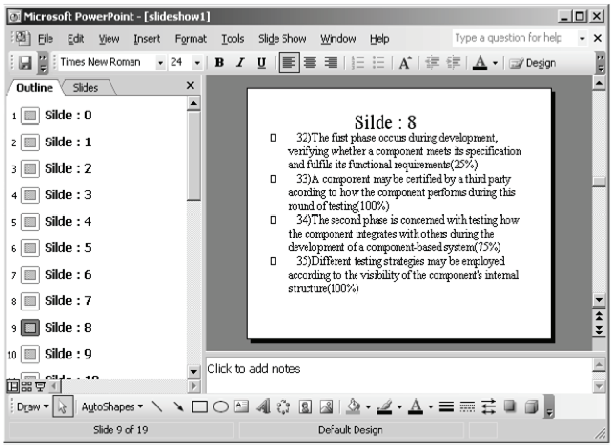
The average precision and recall for user1 for the topic CBSD is 0.8 and 0.88. The average precision and recall for user2 for the topic CBSD is 0.85 and 0.79. The PPT generated by the users in their order of interests for the CBSD topic are shown in the below Fig.17 and 18.



Silde : 0

□ #CBSD ▮ An Overview
□ 1) The development process is briefly covered and some active areas of research in CBSD are described(0%)
□ 2)Component-based software development (CBSD) or component-based software engineering (CBSE) is concerned with the assembly of pre-existing software components into larger pieces of software(100%)
□ 3) Underlying this process is the notion that software components are written in such a way that they provide functions common to many different systems(33%)

Fig.17. PPT generated by user1 for the CBSD topic



Silde : 8

□ 32)The first phase occurs during development, verifying whether a component meets its specification and fulfils its functional requirements(25%)
□ 33)A component may be certified by a third party acording to how the component performs during this round of testing(100%)
□ 34)The second phase is concerned with testing how the component integrates with others during the development of a component-based system(75%)
□ 35)Different testing strategies may be employed according to the visibility of the component's internal structure(100%)

Fig.18. PPT generated by user2 for the CBSD topic

5. CONCLUSION

The presentation is generated for the textual data that are present in the documents. The images are not present in the documents which are taken for the generation of PPT. In future, the presentation can be enhanced by embedding the images, animations and sound.

REFERENCES

[1] Robert M. Losee, June 1998, "Comparing Boolean and Probabilistic Information Retrieval Systems Across Queries and Disciplines". *Journal of the American Society for Information Science*

[2] Manning, Raghavan, and Schutze, 2007, "Introduction to Information Retrieval", *Cambridge University Press*, pp.1-16, pp.169-182

[3] Kuang-hua Chen and Hsin-Hsi Chen, 1995, "A Corpus Corpus Based Approach to Text Partition", *Proceedings of the Workshop of Recent Advances in Natural Language Processing,* pp. 152-161

[4] Han-joon Kima, Sang-goo Lee, 2003, "Building topic hierarchy based on fuzzy relations", *Elsevier publications of Neuron computing*, pp.481-486

[5] Youngjoong Ko, Jinwoo Park, Jungyun Seo, 2004, "Improving text categorization using the importance of sentences", *Elsevier publications of an Information Processing and Management*, pp.65-79.

[6] White, R. W., Jose, J. M., Ruthven, 2005, "Using top ranking sentences to facilitate effective information access", *JASIST*, Vol. 56, No. 10, pp.1113–1125

[7] Hyo-Jung Oh, Sung Hyon Myaeng, Myung-Gil Jang, 2007, "Semantic passage segmentation based on sentence topics for question answering", *Elsevier publications of Information Sciences*, pp.3696-3717.

[8] Shailendra Singh, Lipika Dey, 2005, "A rough-fuzzy document grading system for customized text information retrieval", *Elsevier publications of Information Processing and Management*, pp.195-216

[9] Xiao luo, A. Nur zincir- Heywood, 2004, **"**Combining Word Based and Word Co-Occurrence Based Sequence Analysis For Text Categorization"**,** *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, pp.1580-1585.

[10] Softpedia, free-to-try software programs, http://www.softpedia.com/get/Others/Home-Education/Wintree.shtml.