ISSN:0975-9646

An Improved Annotation based Summary Generation for Unstructured Data

Teena Bhawsar[1] , Devendra singh Kaushal[2]

[1]Department of Computer science & Engg,
Jawaharlal Institute of Technology, Borawan,
Khargone (M.P), India
[2]Assitant Professor,Department of Computer science & Engg.,
Jawaharlal Institute of Technology, Borawan,
Khargone (M.P), India

***Abstract*— In today's era digital documents is facilitating its users towards ease of access and reading of the content which is made available online. The designer emphasizes further enhancements in supporting features to edit or customize the documents according to their needs. All it needs to extract the content from several sources and change them according to user's habits. If the sources and the outcomes belong to some online medium then it comes under the area of web mining. Annotation is one of such process in which the content according to user passed keywords or paraphrases are focused or heighted using software applications. They will not only annotate but also make the collaborative document authorized. The majority of people read and annotate daily, but do not create new documents. With this shift in perspective, there is an increased focus on software primarily targeting reading and annotating. The reading-centric products are aware of the importance of pagination over scrolling, of side margins, and of the relationships between font size, line spacing, and line width.**

But there is a problem associated with current annotation process is that it cannot maintain the modifications because of frequent update of digital documents. The process starts with identifying the user looked keywords in the document and then adding some additional information about their use and description. It comes under the process of knowledge discovery in texts (KDT) or text data mining [1]. In KDT usually plain textual documents are used. There are also some minor attempts to use (partially or fully) structured textual documents as HTML or XML documents in order to make use not only of plain textual parts but also of additional structural information. In this work we have not only proposed the approach but also evaluate its results, that gives the idea that initial results of the approach are satisfactory.

***Keywords*— Web Content Mining, Text Summarization, LSCA, Feedback, Tags, Classes, Rules, Annotations, Precision, Recall, F-Measure;**

I. INTRODUCTION

World Wide Web (WWW) is the popular interactive medium holding massive amount of information and data openly available on internet for its users. The data available online is the collection of documents such as xml, databases, audios, video, text, html etc. The data which is not indexed and are presents in the file formats then it is known as unstructured data and the data which is arranged are known as structured data. Web mining deals with extracting the information or data after processing the user queries to make its access effective and easier. Web mining uses the mining techniques to automatically discover web documents, extract information from web resources and uncover general patterns on the web [2].

Recent areas of work in this field can be separated into two majorly classified domains: mining and retrieval. The retrieval focuses on retrieving appropriate information from bulky repository whereas mining research emphasizes on extracting new information from already existing data [3]. There is a clear separation made between information extraction which focuses on extracting relevant facts and information retrieval focus selects relevant document. Now, Web mining is a part of both information extraction and information retrieval and it also supports the machine learning activities which improves the text classification [4]. The different types of web mining approaches are shown in figure 1.

Web mining is integration of information that is gathered by traditional data mining techniques with information gathered over World Wide Web. It is decomposed into following subtasks [5]:

a) ***Resource Discovery:*** *It helps in retrieving services and unfamiliar documents on web.*
b) ***Information selection and pre-processing****: It automatically selects and pre-processes specific information from the web sources.*
c) ***Generalization:*** *It uncovers general pattern at individual web sites as well as across multiple sites.*
d) ***Analysis:*** *It validates and interprets the mined pattern.*
e) ***Visualization****: It presents the result in visual and easy to understand way.*

Web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining.

Web Content Mining

Web content mining is the sub area of web mining which involves analysis and extraction of text, videos, graphs and pictures based on users query from web sources. It could be further divided into two primary category i.e. agent based and database approach. The agent based approach applies relevant searches of information and generates an organized content. The later one helps in applying retrieval from semi-structured or structured data present online.

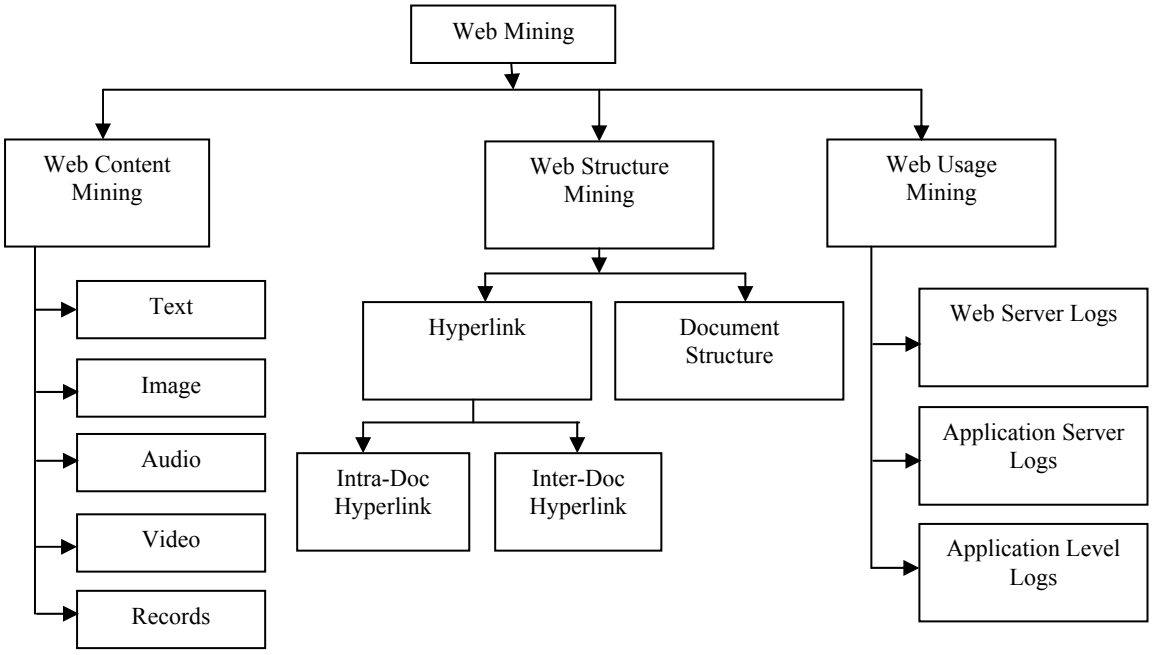


Figure 1: Web Mining Approaches

Their main task is to analyses the content from heterogeneous web resources of web pages and documents designed to help the retrieval for users. Thus its primary goal is to improve filtering and finding of the information as demanded by the users through queries. According to their operations, behavior and usages, web mining is defined in following types

- ***Unstructured Data Mining***
  Here the mining is applied to unstructured information having major problem associated is their massive size. The research of applying these techniques over such a larger area comes under the categories of knowledge discovery in texts [6]. Some of its areas include information extraction, topic analysis, summarization, clustering and visualizations.
- ***Structured Data Mining***
  Here the data is organized in defined structure which facilitates the extraction process. Mainly the extraction is performed from the web pages and organized in the form of list, tables or tree. Some of its well known applications are: page content mining, web crawler, wrappers etc.
- ***Semi-Structured Data Mining***
  In this process the task of source is specifically defined to prevent the structures settings on data. It can extract the data from web and add it into the existing databases. Examples of this approach are: web data extraction language, object exchange model, top down extraction etc.
- ***Multimedia Data Mining***
  This process analyses and finds interesting or related patterns from the media data stores having collection of videos, audios, images and texts using filtered queries.

II. BACKGROUND

Information Retrieval and Annotations

Information retrieval is the approach to extract the information from heterogeneous sources after applying the process of indexing, clustering, classification, filtering and retrieval. It identifies the relevant information from collection of data sources based on metadata or content types. The process starts with object or entity that is represented by content collection of databases and the user queries are matched from these information load databases. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked [7]. This ranking of results is a key difference of information retrieval searching compared to database searching. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the queries.

Annotation is the one of the major operation of information retrieval. Here the additional information is passed as a comment, explanation, notes or other type of remarks in selected part of document to make it more readable by the users. It was an external entity thus, annotation does not require document to be edited. These are stored on specific annotation servers with following properties:

- The location of annotation storage must be either physical or in an annotation server.
- The scopes of annotations are associated with complete document or on small fragments only.

- It can be applied by using comments, annotations and queries.

Text annotations are the type of metadata which includes adding extra notes written for user's private purposes. It can be also applied along with the shared annotations written for the purposes of collaborative writing, editing, social reading, sharing etc. In some fields, text annotation is comparable to metadata insofar as it is added post hoc and provides information about a text without fundamentally altering that original text. Text annotations had identified four functions to serve the requirements of digital documents. These are:

- It facilitates the reading and writing operations using annotations for professional and personal purposes.
- It can perform the sharing of annotations with common search queries.
- It provides effective feedback on applied annotations for collaborative access.
- It emphasises on important topic of the document using footnotes and other functions.

Annotations & Context

A set of sentences is extracted from a given document. The annotation process identifies the sentence location along with its context. Each sentence is made of group of keywords which are extracted from these given annotations and their contexts as metadata. It ranges from one to several sentences and even one sentence may include several annotations, and an annotation may contain several keywords. For each sentence, the keywords are extracted and further identify the annotations it contains. Such sentences are called as annotated sentences. Similarly the keywords occurring in annotations are called annotated keywords. While the keywords occurring in annotated sentences are also called as context keywords.

- ***Keywords Extraction***
  From the document, content words are extracted by counting the frequencies are beyond a certain threshold and not occurring in stopping wordlist. Word frequencies are calculated. After applying word occurrences statistics to full text is generated which somewhere shows the importance of that keyword in the given document?
- ***Sentence Extraction***
  Sentences are weighted according to the keywords it contains which shows the length of a sentence by counting the total number of keyword it is holding. Sentences are ranked by their weights, and then top scored sentences are selected as important ones and used to compose into a summary according to their original position.

Annotation Model

There are several models for adding annotations to documents. In this section, we will discuss three of them. The first possibility is to add metadata to documents without relating the metadata to the document content or parts. XMP, for example, uses this method. The second modeling alternative is to relate the metadata to sections of the document text and other document parts. The advantage of the latter model is that it enables tight integration between documents and ontologies. For example, the model enables users and application programs to use the document text to look up parts of the ontology and vice versa .

Finally, it is possible to store metadata outside the documents, for instance in a separate meta-level database. The advantage of this approach is that no changes are required to the documents. However, the metadata do not follow the documents if they are copied, moved, or communicated to others electronically. Moreover, it is not possible to collect metadata from documents published on the web [8].

Whether a term is ontological is a social matter and not a technical or formal matter. It is sometimes mistakenly understood that using a formal ontology language makes terms ontological. Ontology however denotes a shared (social) understanding; the ontology language can be used to formally capture that understanding, but does not preclude reaching an understanding in the first place. Summarizing, we can distinguish three types of annotations: 1. informal annotations, 2. formal annotations, that have formally defined constituents and are thus machine-readable, and 3. ontological annotations, that have formally defined constituents and use only ontological terms that are socially accepted and understood [9].

Semantic Annotation

Semantic annotation is the process of inserting tags in a document to assign semantics to text fragments allows creation of the documents which processes not only by humans but also automated agents. However, considering the scale and dynamics of worldwide web, application of the traditional natural language processing techniques to annotate documents semantically must be revised. From the engineering perspective there is a number of requirements important to be faced when designing a text processing system:

(i) ***Accuracy:*** *performance must be estimated to access the ability of the tool to retrieve all and only correct answers;*

(ii) ***Flexibility and robustness:*** *these features characterize the viability of a system under abnormal conditions and stability to different text types or domains;*

(iii) ***Scalability****: space and run time limitations must be overcome;*

(iv) ***Data sparseness:*** *dependence on expensive training resources can be an obstacle for porting the tool in a different domain;*

(v) ***Complexity:*** *long response time can render a system unacceptable for human users;*

(vi) ***Multilingualism****: independence from character encodings, lexicographic sorting orders, display of numbers, dates etc. needs to be ensured.*

III. LITERATURE REVIEWS

In the last few years there are several approaches designed and implemented for improving the traditional structure of overall annotation process. Among them we have made a study for analyzing their problems and

solutions which somewhere affects the annotations. Let us start with the brief overview and proceeds towards the detailed algorithmic details available with the paper.

The paper [10] emphasizes on primary understanding about the automatic text classification approach applied using machine learning. They deal with automation of extraction of information based on relativity index of all the documents. The process of text classification leads towards automatic extraction and filtering of keywords and paraphrases. The paper shows an implementation review on the machine learning based prototypic tool. The paper also presented with a detailed survey on various approaches of the text classification like Incremental text classification, multi-topic text classification, discovering the presence and contextual use of newly evolving terms on blogs etc. are some of the areas where future research in automatic text classification can be directed.

In the paper [11], two approaches are presented and evaluated for applying annotations in linked open data sets (LODS). The algorithms are using word sense disambiguation mechanism which uses relationship between the resources and other redefines the definitions presents in the datasets. The applicability of the approaches is also tested on WordNet, Dbpedia and OpenCyc annotating tools. While executing the algorithms the paper also finds the issues regarding the over fitting of datasets. Thus the direction also suggests the use of LODS for further improvements.

Annotation can be applied manually or automatically based on the users requirements and ease of applications. An automatic annotation for documents segments with rich text and domain ontologies are given in *[12]*. The work mainly uses the input document and then extracts its logical structure in different informative units. It has made an assumption that the documents segments must be organized in a hierarchical manner with informal ontologies for creating the meta-data labels or tags. The results of carrying out these experiments demonstrate that the proposed approach is capable of automatically annotating segments with concepts that describe a segment's content with a high degree of accuracy.

The paper [13] worked on applying the annotations on web documents without editing it by some external process. They are applied as a meta-data which holds the additional information about the data and the structure of that document inserted in unstructured text. It can also works for bigger sized data used in organizational information exchanges with collection of unstructured elements. Thus to solve this issues information retrieval approaches are used for extracting the relevant information. There are number of techniques which are useful for obtaining best annotation for documents. Techniques contain extracting information from raw data, extraction of structured metadata and many more.

For organizational purposes the solution to annotation mechanisms are dynamically changes based on their needs and policies. They are having larger size unstructured data containing structured information in it. Traditional information extraction algorithms will only facilitate the relationships identification in very expensive and inaccurate manner.

Thus the paper [14] presents the novel approach for detecting the relativity based information extraction using querying databases. This system works on human interventions towards annotation process which was assume to be more specific and accurate. The paper is given with the algorithms that identify structured attributes that are likely to appear within the document, by jointly utilizing the content of the text and the query workload. Our experimental evaluation shows that our approach generates superior results compared to approaches that rely only on the textual content or only on the query workload, to identify attributes of interest.

Carrying forward the above work, *the paper [15]* gives a solution for web content mining based annotation for dynamic environment. It covers the various files types available for extracting the information from them such as HTML, XML, multimedia, pictures and others. The paper also deals with the problem of information explosion for effectively applying the relevance based tags in documents. The paper uses extraction mechanism which fetches the content from web pages based on user's passed queries. Majorly the article emphasizes the use s of classification and clustering for detecting phishing websites.

In the paper [16], extraction or retrieval is performed on web sources such as forums, blogs, and news articles. Such sources are having high heterogeneity and complexity associated with mapping and extraction of information's. Even these sources are having frequent modifications in their data and their nature. Thus the paper focuses on implementing the homogeneous solution to this heterogeneous problem using indexing techniques for web sources. They holds the necessary information associated with those files into some other files from which metadata is easily maintained and passed to annotation modules. Also the automatic indexing based approach suggested by the papers used for semi structured and unstructured documents are applied in collaboration with MapReduce programming model. Experiments on a real-world corpus show that our approach achieves a good performance.

The paper [17] presents the detailed survey of various web mining approaches for analyzing the content available with different web sites. It opens the current work and opens the new possibilities for getting the better accuracy and high reliability towards content extractions and mining tasks. Majorly the suggested methods worked on measuring the relevance values for the documents based on contents and the passed queries. Overall organization of the paper also covers the details related with web content mining and their approaches. It also covers the application of these approaches for structured, unstructured, semi-structured and multimedia data mining techniques.

Another survey of web mining based approaches for text analytics is given *with the paper [18].* It applies semantic analysis on different types of documents for getting the improvements in annotation based search. Although there are various techniques implemented for the efficient searching of using annotations. Here in this paper a survey and analysis of various annotations based

techniques are analyzed and discussed here so that on the basis of their various advantages and limitations a new and efficient technique is implemented in future.

IV. PROBLEM IDENTIFICATION

Web content mining is the approach used to extract the content from heterogeneous sources into single output that is more useful to user. It applies mining, extraction, integration of different knowledge entities into single units. Mostly the issues coming in this area is regarding the nature of semantic data and the desired results. Also while getting the annotated text summarization there are some content regarding queries which was not effectively resolved. While manually annotating the documents what we do, we analyses the text before making the annotation and separates the words that are not clearly defined or leave those unfamiliar terminologies. After in depth analysis of previously developed approaches related with text summarization and annotation we have found some problems which was not effectively resolved.

- Decision of importance of sentences based on their keywords must follows some rules apart from just counting the frequencies count of words.
- Dealing with stop list words, ambiguity and noisy data are not handled properly. Also with ambiguous words it was very complex to generate the summary and annotate the document.
- Semantically and grammatical analysis must be strong enough to remove the inappropriate sentence compositions and also the data dependencies must be removed first before summarization.

After analyzing the above problem areas, we tries to overcome the above issues by dealing with each problem individually which somewhere affects the annotated summary generation process. Here the aim is to increase the accuracy of summary relativity with the annotations applied. As a direction we also extract the importance of information which was somewhere presents as content or meta-data of that file. Removal of noise based data automatically improves the summarization accuracy.

V. PROPOSED SOLUTION

Web miming is the recent area of work in which the effective content is extracted from the heterogeneous sources based on their characteristics. This work will emphasize on applying the annotation based on content attributes which compares itself with the passed queries. With only annotation the work seems to be incomplete because the content passed by the user might not be related to the desired content and if the quality is matched with bunch of document then the selected document may mislead the directions. Thus after analyzing all the documents then only the annotation can help in reducing the reading efforts.

With this work the intension is to generate the annotated summary coming from the variable content types like from unstructured and semi-structured sources. Summary generation comes under the text summarization process of web mining which could be extractive or abstractive. If we are using the sentences and terminologies which was already there then it is an extractive process else it is an abstractive process.

Important Components of the solutions:

- Segmentation Module
- Latent Semantic Content Analyzer (LSCA)
- Rule Repository
- Singular Value Decomposition (SVD) Module
- Feedback Mechanism based on Quality of Extraction
- Tag Generator
- Annotation Based Summary Generator
- Result analyzer

The process starts with passing the document to the system in docx or text format only as we are focusing on specific file types. These documents are segmented on the basis of their types and content holding probability and the weight age of the content keywords. We break the documents into smaller part which can be easily processed by the system. Later on these decomposed elements are passed onto the latent semantic content analyzer (LSCA) module. It work is to identifies the statistical and linguistic features from the sentences. They separates the keywords and their frequency based on language rules and let them store to the LSCA repository. Now both the features are simultaneously passed to two distinct modules.

Here the first one emphasizes on summary generation and the second one is for classes formation based on content types. For generating the summary SVD (Singular Value Decomposition) is used which assigns the values to each sentences based on their frequencies and content relativity. It will generate the similarity matrices based on the content quality and relatedness. It is of two types: statistical and semantic while its unit is words, phrases, vectors and hierarchies. This module maintains the relation between the content, its axioms and the instances of each segment. Also the sentences containing noisy or corrupted data, repeated keywords, stop words larger than the total keywords of those sentences, ambiguous words not mapped with dictionaries, and dependent statements are separated from the other content.

They are further analyzed by the SVD modules and works after the first set of results came from other part of given input. Once the first phase was over with direct sentences then these problematic sentences are passed again by taking their inputs results and feedback of first phase. Thus this noise and ambiguity removal is an evolutionary approach works towards improving the quality of inserted text.

Now the sentences having the larger similarity values are selected for generating the summary. The second module works for class formation based on the extraction rules. Once the classes are generated then the tags based on them are filtered out. The annotated summary generator combines the selected sentences based on relativity and applies the class tags on it to generate the annotation.

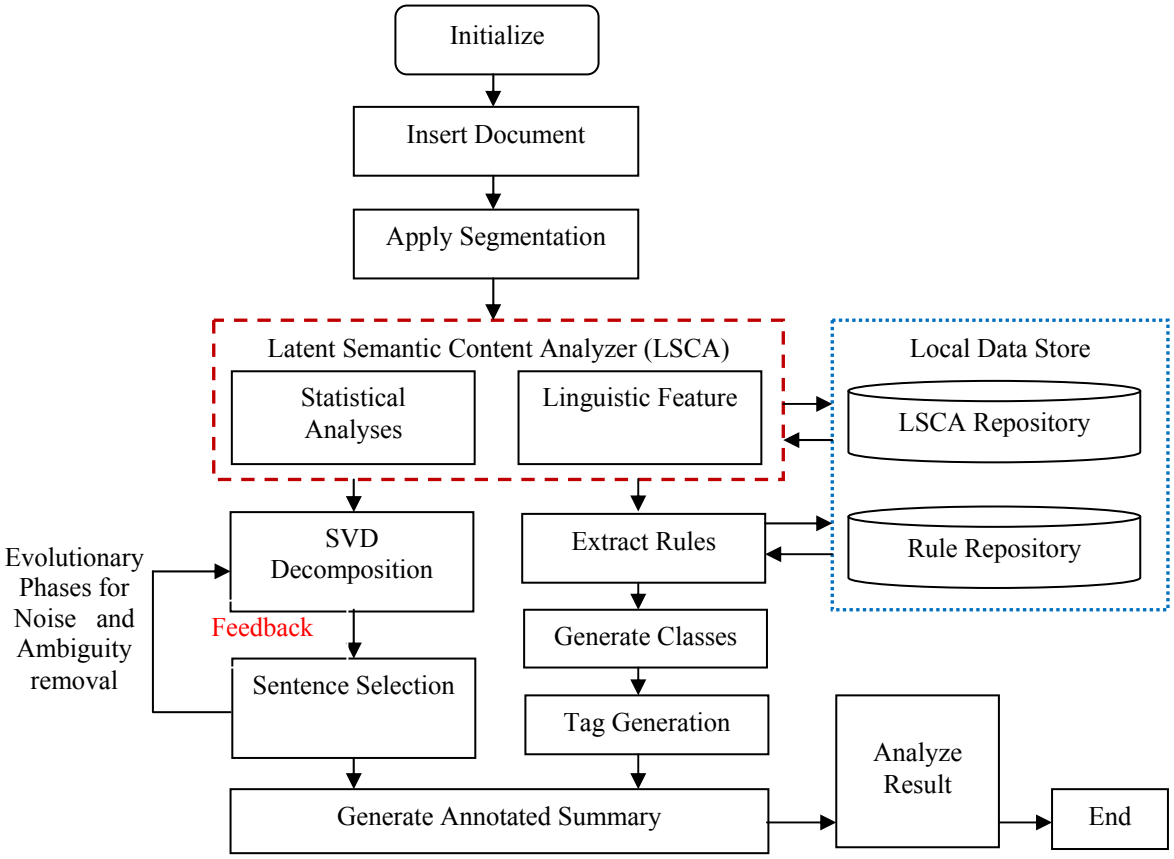


Figure 2: Proposed Annotation Based Summary Generation

The overall process is monitored to get the parametric result evaluation for comparing the outcomes with the traditional approaches. The system maintains the local data repository for tags, classes, rules, semantic features and linguistic analysis.

VI. Evaluation Parameters

For evaluation, a comparison has to be made between human-annotation and generic annotation given by the system. There are a lot of measures to make the comparisons such as precision, recall, some of which will be used for our evaluation. Recall is a measure of how well the tool performs in finding relevant items, while precision indicates how well the tool performs in not returning irrelevant items. In this evaluation, we also took into account partial answers, giving them a half score, as shown in formulas (1) and (2). The annotations are *partially correct* if the entity type is correct and the contents are overlapping but not identical.

Recall = (TP + ½ Partially Correct) / (TP + ½ Partially Correct + FN)

Where TP – true positive answers, FN – false negative answers (1)

Precision = (TP + ½ Partially Correct) / (TP + ½ Partially Correct + FP)

Where TP – true positive answers, FP – false positive answers

Precision and recall are generally applied to sentences; in fact they can be applied to keywords too, which reflects the percentage of keywords correctly identified. Therefore, in spite of summary similarity, our measures for evaluation also include sentences precision, sentences recall, keywords precision and keywords recall. For sentences evaluation, a sentence annotation is correct if it has as many possible keywords as in the corresponding sentence in the human-made summary, that is, their similarity (calculated same as summary similarity) is beyond a certain threshold.

We also calculated F-measure, the harmonic mean of recall and precision:

F−measure=[2×Recall×Precision]/Recall + Precision

This is also known as the F1 measure, because recall and precision are evenly weighted.

VII. Result Evaluation

For evaluation, a comparison has to be made between human-annotation and generic annotation given by the system. There are a lot of measures to make the comparisons such as precision, recall, some of which will be used for our evaluation. To determine how to captures the correctness of the result we need to form a confusion matrix U into the sets and for that use several approaches.

Actual / Predicted	Negative	Positive
Negative	a (FN)	b (TN)
Positive	c (FP)	d (TP)

Table 5.1: 2*2 confusion matrix

Recall is a measure of how well the tool performs in finding relevant items, while precision shows how well the tool performs in not returning irrelevant items. In this evaluation, we also took into account partial answers, giving them a half result, as indicates in formulas (1) and (2). The annotations are *partially correct* if the entity type is correct and the contents are overlapping but not identical.

Precision and recall are generally applied to sentences; in fact they can be applied to keywords too, which reflects the percentage of keywords correctly identified. Therefore, in spite of summary similarity, our measures for evaluation also include sentences precision, sentences recall, keywords precision and keywords recall. For sentences evaluation, a sentence annotation is correct if it has as many possible keywords as in the corresponding sentence in the human-made summary, that is, their similarity (calculated same as summary similarity) is beyond a certain threshold.

Performance measures used to evaluate these algorithms have their root in machine learning. A commonly used measure is accuracy, the fraction of correct recommendations to total possible recommendations.

$$Accuracy = \frac{(correct\ recommendations)}{(total\ possible\ recommendations)}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Where N = TP+FP+TN+FN is the total number of items which can be recommended.

$$Precision = \frac{(correctly\ recommended\ items)}{(total\ recommended\ items)}$$

$$Precision = \frac{(TP)}{(TP + FP)}$$

$$Recall = \frac{(correctly\ recommended\ items)}{(total\ useful\ recommendations)}$$

$$Recall = \frac{(TP)}{(TP + TN)}$$

To find an optimal trade-off between precision and recall a single-valued measure like the F-measure can be used. The parameter α is controls the trade-off between precision and recall.

A popular single-valued measure is the F-measure. It is defined as the harmonic mean of precision and recall.

$$F - Measure = \frac{2 * Presicion * Recall}{Precision + Recall}$$

$$F - Measure = \frac{2}{\left(\frac{1}{Precision}\right) + \left(\frac{1}{Recall}\right)}$$

It is a special case of the E-measure with α = .5 which places the same weight on both, precision and recall. In the LSCA summary generation based annotation process the F-measure is often referred to as the measure F1.

Result Interpretation: The above table covers the basic details of complete process applied for generating the summary. Here the time based measurement is performed to analyze the tools behavior that how fast the result is generated. While closely looking at table it is found that he generation time is reduced far files having size in between 5 to 50 kb.

While taking it as in words limits the system is successfully applying the process in between 500 to 1000 words. For generating the summary larger than this word limit the system is considering some additional delays. It can be taken as future work to overcome this. Now as far as the accuracy of the system is considered we need to capture some other parameters like accuracy, precision, recall and F-asure.

Table 5.2: Evaluation of Performance Procedures for Proposed Tool

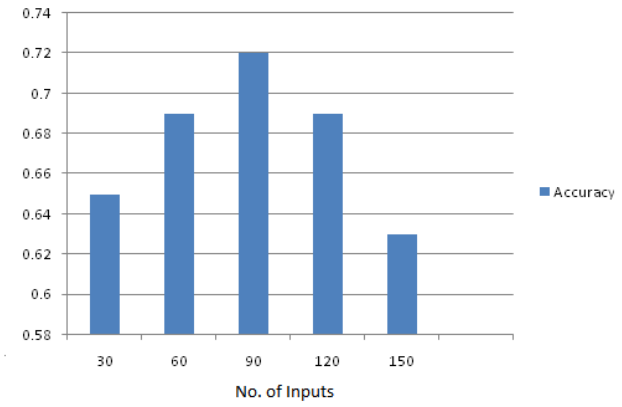
S. No	Input Type	No. of Items	Time Based Evaluation (ms)				
			Segmentation	Rule Extraction	Class Generation	Tag Generation	Summary Generation
1	Positive	3916	300	5955	3178	2987	1597
2	Neutral	869	114	2071	1651	1466	1746
3	Neutral	683	77	2336	1229	1603	1473
4	Positive	279	148	4336	1153	1337	6168
5	Neutral	495	32	3076	1411	1566	858
6	Neutral	869	132	2320	1740	1505	1127

Table 5.3: Evaluation of Proposed LCSA Annotation Based Summary Generation

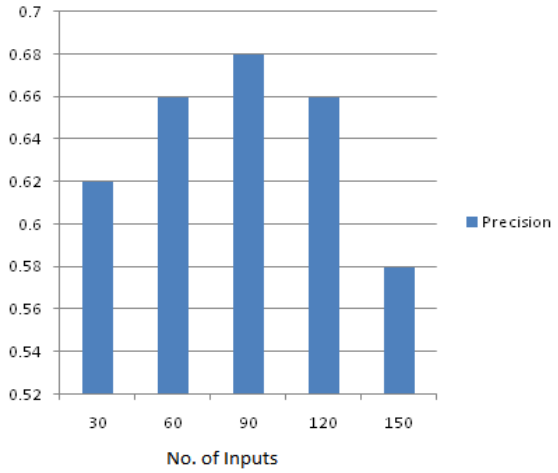
S. No	Input Type	Sentiments	No. of Inputs	Accuracy	Precision	Recall	F-Measure
1	Text (File)	Positive	150	0.63	0.58	0.68	0.68
2		Neutral	120	0.69	0.66	0.73	0.73
3		Positive	90	0.72	0.68	0.69	0.69
4		Positive	60	0.69	0.66	0.72	0.72
5		Neutral	30	0.65	0.62	0.77	0.77

Result Interpretation: The above table shows the comparison between the various factors of the recall, precision, accuracy & F-Measure. After analyzing the various captured values for different execution sets the result table shows the behavior of developed project. Here the value of accuracy is continuously gets increased for various detection of passed statements. There is very change observed in the value of precision, recall & F-Measure. By the above factors and their observed values it is clear that the system is capable of detecting the polarity of the post and annotates the statement accordingly.

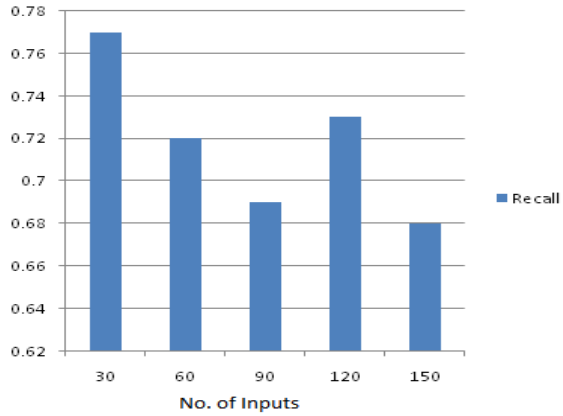
Graph Based Analysis



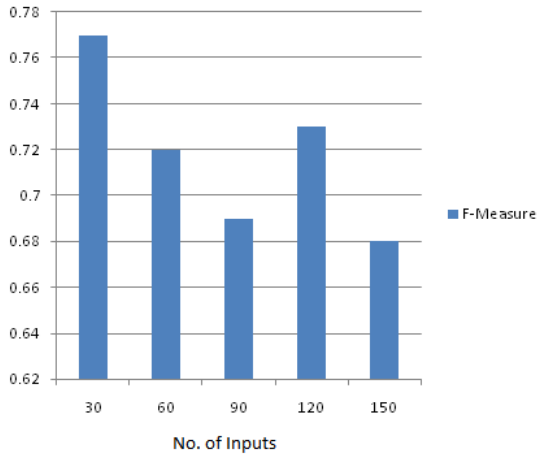
Graph 1:No. of inputs and Accuracy.



Graph 2: **No. of inputs and Precision.**



Graph 3: **No. of inputs and Recall.**



Graph 4: **No. of inputs and F-Measure.**

The above graphs are used for visualization of the tool behavior. We can compare the result by some existing summary generation or semantic analyzer system for different types of the system. Here as we have no executions of previously developed system with their exact dataset or the input we are separately taking the results for our own system. From the above graph it is clear how the value of accuracy, precision, recall and F-Measure is varying according to the different input sets. All these values are dependent on the calculation made for TP, FP, TN & FN. Thus by final valuation process we can say that the tool is outperforming its competitors by analytically comparing the values obtained by our tool.

VIII. BENEFITS OF AUTOMATED ANNOTATIONS

Annotations that are not immediately picked up by automated processing systems are still useful: users can always read them directly, and mining technology is likely to improve. Authors should not be prevented from entering precisely the content they think is needed simply because the system does not know how to use it yet.

(i) It helps tremendously in the analysis and synthesis of information.

(ii) One of the mayor benefits of annotation is context. It is the comment in relation/on top of the data that makes annotation a powerful resource.

(iii) With Annotations you are not making any changes to the actually document. You can add a note or highlight something; however you will not be able to add a sentence.

(iv) Security is also an additional benefit as you can track who made annotations and you can prevent people from making actual changes to the document.

(v) This kind of annotation leads to the better precision of the information retrieval process - by expressing the context of searching in a more precise manner.

Moreover, such more expressive form of describing content of information resources supports a more powerful knowledge sharing process enabling the discovery of new information by considering the combination of existing information resources. In that way, the system provides some answers which are not explicitly stated in the information repository. We present a knowledge management framework, which implements such type of annotation and give a small evaluation study. The framework is supported by Semantic Web technologies, which are based on the machine-understandable description of document content, enabling in that way the automation of the knowledge sharing process.

IX. CONCLUSION

Data is extensively stored in the wide variety inside the data center. Data and information is stored in text files along with the un structured data as well. Flat files are very popular means to store information in Microsoft word format. Annotation is the very efficient way to get the summarized view of information. With this work the intension is to generate the annotated summary coming from the variable content types like from unstructured and semi-structured sources. Summary generation comes under the text summarization process of web mining which could be extractive or abstractive. If we are using the sentences and terminologies which was already there then it is an extractive process else it is an abstractive process. In this work we first one emphasizes on summary generation and the second one is for classes formation based on content types. In this work we evaluated the results produced by the implemented prototype and was satisfactory at the initial phase of research .

REFERENCES

[1] A.J. Bernheim Brush, David Bargeron, Anoop Gupta, and JJ Cadiz " Robust Annotation Positioning in Digital Documents" in Microsoft Research 2000.

[2] Jan Paralic and Peter Bednar" Text Mining for Documents Annotation and Ontology Support " in Web Technologies Supporting Direct Participation in Democratic Processes 2001

[3] Anne Kao & Steve Poteet " Text Mining and Natural Language Processing –Introduction for the Special Issue" in SIGKDD 2005.

[4] Nadzeya Kiyavitskaya, Nicola Zeni,James R. Cord, Luisa Mich and John Mylopoulos " Semi-Automatic Semantic Annotations for Web Documents"

[5] Marius Pa, Benjamin Van Durme, Nikesh Garera " The Role of Documents vs. Queries in Extracting Class Attributes from Text" in ACM 2007.

[6] Henrik Eriksson "An Annotation Tool for Semantic Documents (System Description)".

[7] S.R.K. Branavan Harr Chen , Jacob Eisenstein , Regina Barzilay " Learning Document-Level Semantic Propertiesfrom Free-Text Annotations" Journal of Artificial Intelligence Research 34 (2009) 569-603

[8] Paolo Ferragina, Ugo Scaiella " TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities)" in ACM 2010.

[9] Rafeeq Al-Hashemi"Text Summarization Extraction System (TSES) Using Extracted Keywords" *International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010*

[10] Mita K. Dalal ,Mukesh A. Zaveri " Automatic Text Classification: A Technical Review" *International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011*

[11] Delia Rusu, Blaž Fortuna, Dunja Mladenić "Automatically Annotating Text with Linked Open Data" LDOW'11, March 29, 2011, Hyderabad, India.

[12] Maryam Hazman, Samhaa R. El-Beltagy and Ahmed Rafea " An Ontology Based Approach for Automatically Annotating Document Segments " IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012

[13] Priyanka C. Ghegade, Vinod S. Wadne"A Survey on Facilitating Document Annotation Techniques " International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 2013

[14] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis "Facilitating Document Annotation Using Content and Querying Value " IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 2, February 2014.

[15] Sobana.E , Muthusankar.D"A Survey: Techniques of an Efficient Search Annotation based on Web Content Mining" International Journal of Computer Applications (0975 – 8887) Volume 104 – No.3, October 2014

[16] Saidi Imene, Nait Bahloul Safia "An Approach for Indexing Web Data Sources" I.J. Information Technology and Computer Science, 2014, 09, 52-58.

[17] Shipra Saini, Hari Mohan Pandey "Review on Web Content Mining Techniques" International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 18, May 2015

[18] Avnish Rajput,Prof. Amit Saxena,Dr. Manish Manoria " Web Mining: A Survey on Various Annotation Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 4029-4032