Circuits and Systems, 2016, 7, 1856-1865
Published Online June 2016 in SciRes. http://www.scirp.org/journal/cs
http://dx.doi.org/10.4236/cs.2016.78160

Content Based Segregation of Pertinent Documents Using Adaptive Progression

Perumal Pitchandi, Sreekrishna Muthukumaravel, Suganya Boopathy

Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, India
Email: perumalp@srec.ac.in, sreekrishna@srec.ac.in, suganyab@srec.ac.in

Received 21 April 2016; accepted 10 May 2016; published 22 June 2016

Abstract

Due to the emerging technology era, today a number of firms share their service/product descriptions. Such a group of information in the textual form has some structured information, which is beneath the unstructured text. A new attainment which facilitates the form of a structured metadata by recognizing documents which are likely to have some type and this information is then used for both segregation and search process. The idea of this advent describes some attributes of a text that will match with the query object which acts as identifier both for segregation as well as for storage and retrieval. An adaptive technique is proposed to deal with relevant attributes to annotate a document by satisfying the users querying needs. The solution for annotation-attribute suggestion problem is not based on the probabilistic model or prediction but it is based on the basic keywords that a user can use to query a database to retrieve a document. Experiment results show that Querying value and Content Value approach is much useful in predicting a tag for a document and thus prediction is also based on Querying value and Content value which greatly improves the utility of shared data which is a drawback in the existing system. This approach is different, as we consider only the basic keywords to be matched with the content of a document. When compared with other approaches in the existing system, Clarity is a primary goal as we expect that the annotator may improve the annotations on process. The discovered tags assist on quest of retrieval as an alternative to bookmarking.

Keywords

Document Annotation, Segregation, Identification, Content Type

1. Introduction

Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intellect, contraption erudition information

How to cite this paper: Pitchandi, P., Muthukumaravel, S. and Boopathy, S. (2016) Content Based Segregation of Pertinent Documents Using Adaptive Progression. *Circuits and Systems*, **7**, 1856-1865. http://dx.doi.org/10.4236/cs.2016.78160

and database system [1]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for foreside from the raw investigation step, it involves Database and data management aspect, data pre-processing, model and inference considerations, interestingness metrics, complexity consideration, post-processing of discovered structures, visualization, and online updating [1] [2]. Many existing organization share their descriptions about products and services. For illustration, Scientific networks, social networks or disaster management group share their information. Prevailing technologies like content management software (e.g.: -Microsoft Share point) allows users to share documents and tag them in a improvised manner Like that, Google Base allows users to define objects for them either by choosing from predefined or to define their own attributes. This process may facilitate subsequent information discovery. Many annotation systems provide a single way for annotation:”un typed” annotation. Consider that a user may annotate a weather report using a tag such as “Storm Category 5” [2]. In general the most effective expressive annotation strategies use ”attribute-value” pairs as they can contain more un typed information than un typed approaches. In such cases, the above information can be entered as (Storm Category, 5). A most recent work in using the most expressive queries is the “pay-as-you-go” querying strategies in Data space in which users provide the data integration hints at the query time [3] [4]. In such hypothesis based system, the structured information is already present and the difficulty is with matching the source attributes with the query attributes.

Some systems don’t have any basic idea about “attribute-value” annotation that makes the “pay-as- you-go” querying feasible. “Attribute-value” pair based annotation requires users to be more focused on their annotation efforts. In such case, the users should know about the underlying architecture and associated field types. They should also be aware of when to use these field types individually. With architecture which requires some hundred’s of information to be filled, the process becomes much complicated and congested. It results in the ignorance of annotation capabilities that is left to be used by the users [5] [6]. Even if some systems allow the users to annotate a document in a random manner. This task requires some effort. The user should have unclear usefulness for subsequent searches in the future—who is going to use an arbitrary, undefined in a common schema, attribute type for future searches. Even if the attribute fields are limited to a particular number, we can’t predict like what fields among them will be utilized to search for searching at the future. Such issues results in naming a document with the very basic keywords. Such simple things make the analysis and querying of a data to be more tedious. In such case, users were often limited with plain text searches or to have access to very basic annotation fields such as “creation date”, ”owner of a document” and so on [7].

In this paper, a cost segregation approach which is similar to CADS (collaborative Adaptive Data Sharing Platform) is proposed. It is an approach which has two ways of segregating a document [8].

1) By “annotate-as-you-create” platform which facilitates fielded data annotation.
2) By “automated annotation” of a document with its content. In contrast, we are to segregate a document based on its content towards to generate attribute values for attributes that are often used by querying users.

CAD Sharing Platform Approach is to stimulate and lower the cost of creating nicely annotated chronicles that can be promptly useful for commonly furnished semi structured queries such as the ones in **Figure 4**. Our key intention stimulate the annotation of the chronicles at creation time, while the originator is still in the document creation stage, despite the fact the techniques can also be used for post generation document annotation. In our framework the originator generates a new document and uploads it to the warehouse. After the upload, CADS examines the text and originates the adaptive insertion form. The form holds the best attribute names chronicle text and information need (query workload), and the most feasible attribute values given the chronicle text. The originator can inspect the form, recast the generated metadata as necessary and acquiesce the annotated chronicle for storage.

2. Related Work

However, for human beings it’s simple to judge whether two words are similar or not. But for a computer, it’s a difficult task, which involves psychology, philosophy, artificial intelligence and other fields of knowledge. Hence a computer being a syntactic machine, it cannot understand the semantics [9]. In order to that semantic associated with words or their meanings are to be represented as syntax. Semantic similarity measurement between words or their meaning is a basic research area in the fields of natural language processing, intelligent retrieval, document clustering, document classification, automatic question answering, word sense disambiguation, machine translation etc. The basic process of data processing is represented in **Figure 1**. Almost all existing studies

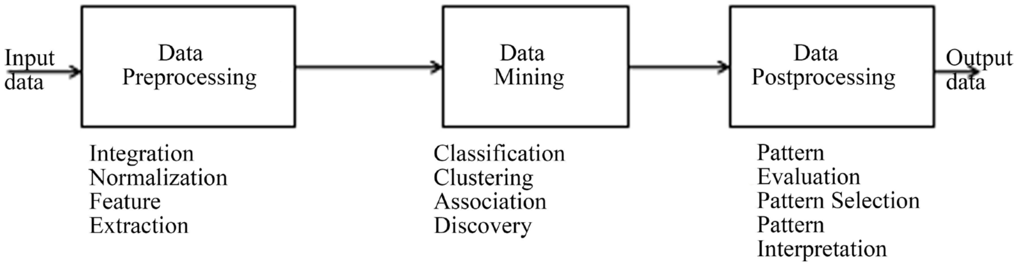


Figure 1. Data processing.

on semantic similarity approaches are only concerned with supervised metrics that are low term coverage and difficult to update [10].

Collaborative annotation Systems like IBM MPEG-7 tool favors this type of annotation for an object and uses the previously used tags for annotations of new objects. An eloquent amount of work has been done to predict tags for documents or resources like web pages, images, videos [13], from the user's perspective and involvement, this approach takes different forms on what is anticipated as an input to the system. However the goals are similar to predict the missing tags that are related to an object. The solution that they have proposed is based on a probabilistic framework that considers the evidences in the document content in the query workload [11]. There are two ways to combine these two pieces of evidence, content value and querying value. A model that considers both the components conditionally independent and a linear weighted model. A technique which suggests attributes that improve the visibility of the documents with respect to the query workload by up to 50% was proposed. The method what they had used to extract the structured information from unstructured text is CADS approach. The CADS approach is the collaborative adaptive data sharing platform which is an "annotate-as-you-create" infrastructure that facilitates field the data annotation. **Figure 1** illustrates the basic steps of data processing.

The consolidated model for CADS is quiet similar to a data space, in which a heterogeneous source is proposed for a loosely integrated model [12]. A cardinal difference is that data space use blending of existing annotations to produce solutions for a query. But our work evinces the appropriate annotations at the insertion time, by considering the query workload to identify the important attributes to add. A real Time application-Google Base A real time application which is a related data model is Google-Base [13], in which users can specify attribute/value pairs of their desire. Information Extraction Information extraction is mainly related with the context of value suggestion for the computed attributes. We can classify the IE into two namely open IE closed IE. Closed IE is much cumbersome but open IE is close to CADS approach. In recent years, for document clustering semantic similarity between words or terms has become an increasingly important research topic in data mining. Semantics identify concepts which allow for the extraction of information from data and looking for the meaning of documents or queries concepts need to be captured. It plays an important role in underlying higher level application and become a key point in research.

3. Limitations in the Existing System

Our inspiring frame work is a disaster management situation, inspired by the experience in building business continuity information network [3] for disaster situations in South Florida. During calamities, we have many users and concerns proclaiming and ingesting information. For example, in a hurricane situation, local government firms report shelter location, damages in structures, or structural warnings. Many algorithms and approaches about semantic similarity measurement between words and their meaning are in the range to improve judgment between documents. The semantic similarity approaches or metrics between words and their meanings can be categorized as supervised metrics and unsupervised metrics. The resource based metrics and knowledge-rich text mining requires such human resources and is referred as supervised metrics. Some problems may rise due to existing algorithms that are low in their efficiency due to their technology lag, tedious task, time consuming and resource restrictions. Because of the vastly available documents and high growth of the document both in size and number, it's difficult to analyze each document separately and directly. It requires lot of human resources.

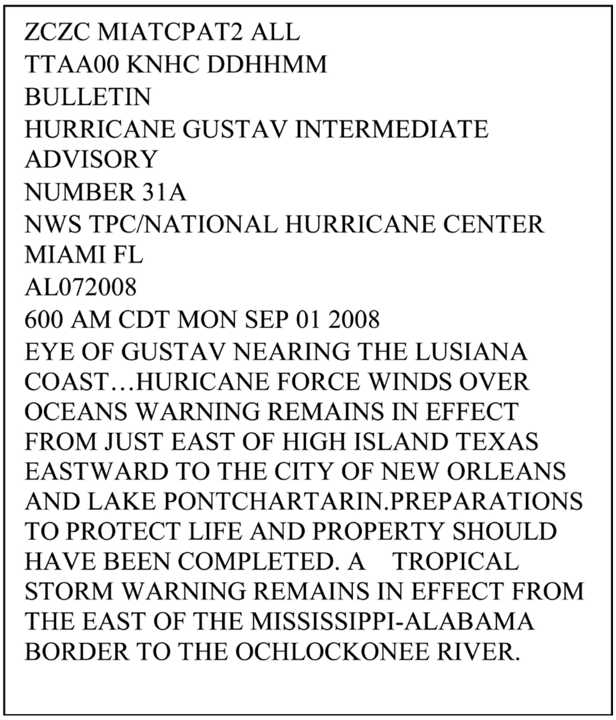
In **Figure 2**, it shows a report extracted from the National Hurricane Center repository, narrating the status of a hurricane event in 2008. The report gives the storm location, wind speed, warnings, category, advisory identifier number, and the date it was revealed. Despite the fact, this is a text chronicle; it contains absolutely many attributes names and values, for example, (storm category).

In **Figure 3** we could improve the standard of the penetrating through the database. For occurrence, **Figure 4** shows three specimen queries for which the report of is a good answer and the lack of the appropriate annotations makes it hard to retrieve it and rank it properly.

The drawbacks of the existing system are the user's interest towards the attribute suggestion is considered at last not at first. To automate this CADS process we need to have a large database. Some users might find the attribute suggested and their values to be useful for storing the document. Some users might find the attribute suggested and their values do not much useful for storing the document. As, humans interest differs from person to person and the importance what they give will also differ.

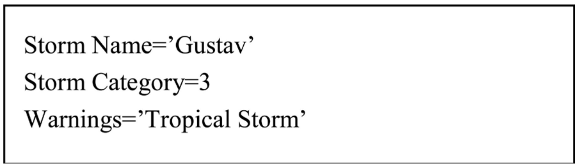
4. Proposed Work

Most recently there has been a huge research interest in developing web based similarity measures. But in this



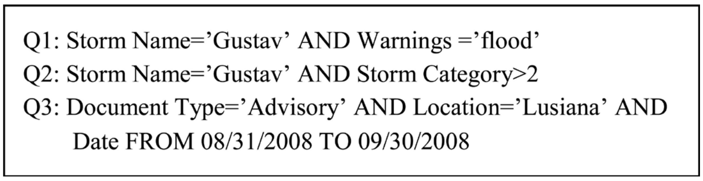
ZCZC MIATCPAT2 ALL
TTAA00 KNHC DDHHMM
BULLETIN
HURRICANE GUSTAV INTERMEDIATE
ADVISORY
NUMBER 31A
NWS TPC/NATIONAL HURRICANE CENTER
MIAMI FL
AL072008
600 AM CDT MON SEP 01 2008
EYE OF GUSTAV NEARING THE LUSIANA
COAST…HURICANE FORCE WINDS OVER
OCEANS WARNING REMAINS IN EFFECT
FROM JUST EAST OF HIGH ISLAND TEXAS
EASTWARD TO THE CITY OF NEW ORLEANS
AND LAKE PONTCHARTARIN.PREPARATIONS
TO PROTECT LIFE AND PROPERTY SHOULD
HAVE BEEN COMPLETED. A TROPICAL
STORM WARNING REMAINS IN EFFECT FROM
THE EAST OF THE MISSISSIPPI-ALABAMA
BORDER TO THE OCHLOCKONEE RIVER.

Figure 2. Example of an unstructured document.



Storm Name='Gustav'
Storm Category=3
Warnings='Tropical Storm'

Figure 3. Desirable annotation for the document.



Q1: Storm Name='Gustav' AND Warnings ='flood'
Q2: Storm Name='Gustav' AND Storm Category>2
Q3: Document Type='Advisory' AND Location='Lusiana' AND
Date FROM 08/31/2008 TO 09/30/2008

Figure 4. Quires that can benefit from the annotations.

work it focus on the problem of fully unsupervised OS based semantic similarity computation between words and their relevant meaning. The proposed algorithm requires no expert knowledge or language resources. Here we investigate the search process which is an unsupervised OS based mapping metrics to find the exact documents that matches the semantic meaning or the exact word. The first is to find the documents that exactly have the given word. The second is fully text based, which classifies a document as T [Good] and T [Best]. In this paper an adaptive technique is proposed for automatically generating segregated data from the whole unstructured documents by annotation such that the utilization of the data when a query is given is maximum. It is created by examining principled probabilistic methods and algorithms to extract keywords from query workload and to integrate those things to form information as a separate document based on those keywords. It deals with real datasets and real users to get the accurate results.

Figure 5 shows the process of the analyzing and segregating the document based upon the users input. One objective is to support planning of the knowledge discovery process and buildings of workflows for a user task. The second objective is to support the meta-mining. The Existing language resources and their algorithms are time consuming and it demands a human resources. To address this problem we propose unsupervised semantic similarity computations between the words, their relevant meanings and terms in the document. This algorithm work automatically and does not require any human supervision. The knowledge for annotation process to save their document is quite not much needed. In addition, the proposed unsupervised context-based similarity computation algorithms are shown to be competitive. With the state of art supervised semantic algorithms that employ language specific knowledge resource and also retrieved useful documents. The advantage of the proposed frameworks is that even if the meaning is known we can get our required documents. If the word is also known we can get the exact documents. Search result is accurate when compared with the results given by the operating systems.

5. Experimental Evaluation

The effect of word search with the existing operating systems is initially investigated for the search results. Basically the measures can be classified in to supervised measure and unsupervised measure. Supervised measure uses hand grafted resources like ontology. Various research journals and picked out several unsupervised

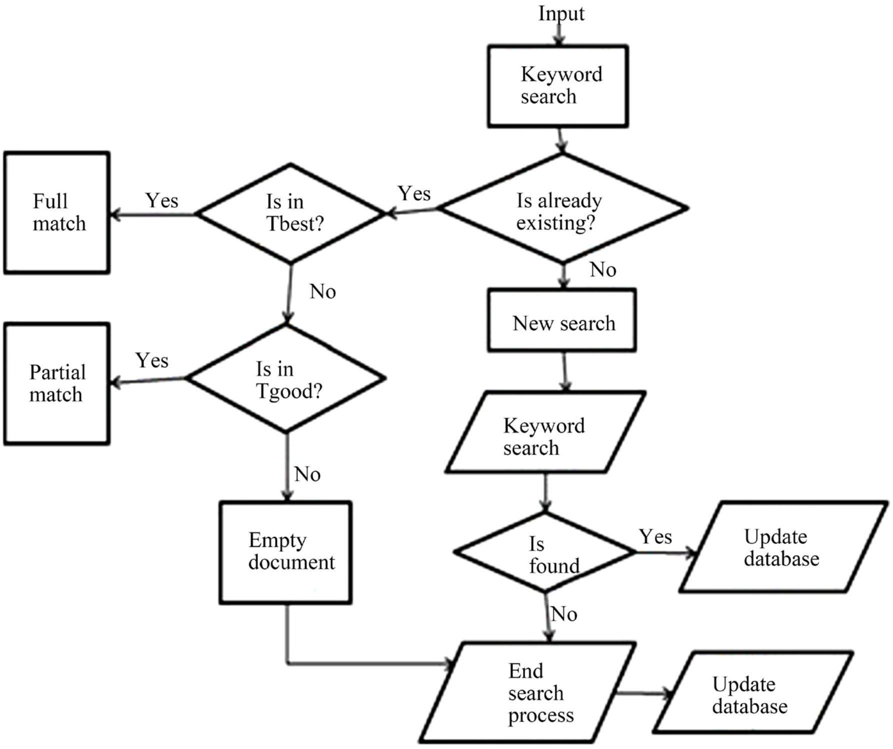


Figure 5. Process of the adaptive approach.

similarity and distance measure that play a vital role in Data Object Clustering. Those similarity measures and highlighting their merits was the mentioned tasks of information retrieval. The best measures which provide efficiency and accuracy in the tasks of information retrieval.

5.1. Attribute Suggestion

The problem that we are going to deal is attribute suggestion problem. To potentially solve this problem we recognize and recommend attribute for a document "d" with two properties querying value (QV) is the collection of basic keywords for a content type of a document. Content value (CV) is the set of keywords from the content of a document (dt) and example is shown in **Table 1**.

Based on the above two properties we process a document we find the keywords in CV which matches with the QV as shown in **Figure 6**. The real play starts here, when a keyword is read in a document it is processed with the QV to find the type of the CV. If exact match is found the next keyword matching with the QV is found from the CV. Maximum keyword count is found from this process and type of the document is thus identified. Then the document is annotated based on the type of the document that is identified by the result with the QV. We first introduce the two optimal suggestion techniques namely,

1) OPT Full Match—It is a technique which uses the subset of the ground truth attributes for each document that satisfies the maximum number of queries. It is an NP-hard problem. For simple workload it works well at the same time for a huge workload it takes some significant amount of measurable time.
2) OPT Partial Match—It is a technique which maximizes the number of query conditions satisfied. It is found by making a single pass on workload.

Table 1. Specifying attributes.

Attribute Name	Attribute Value
Storm Name	Gustav
Storm Category	3
Warnings	Tropical Storm
Storm Speed	16mph
Location	Louisiana
Max Wind Speed	115mph

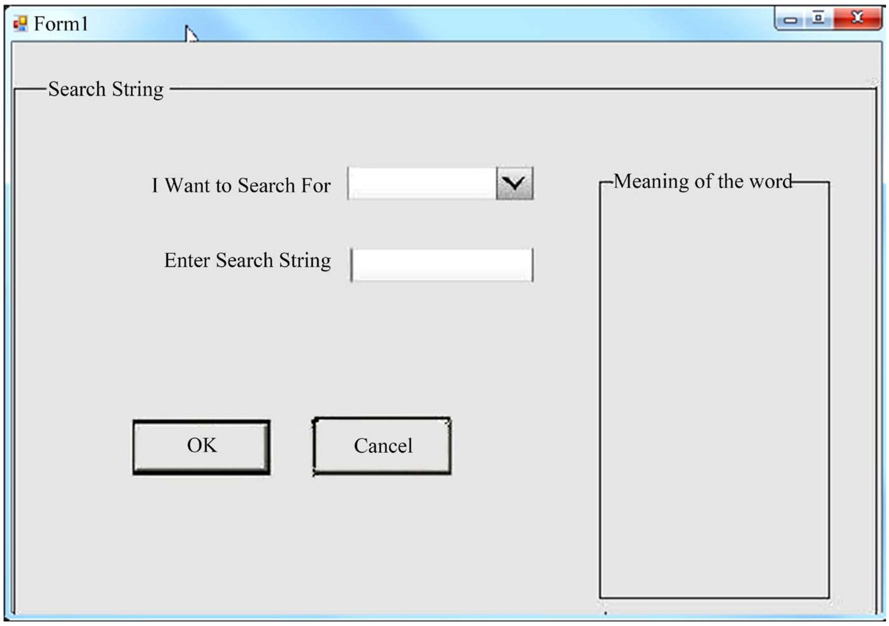


Figure 6. Provide the string for searching.

5.2. Keyword Fetching and Matching

Initially, the user has to specify his interest as either he is going to search with Meaning or Word as a string. If he doesn't know the exact word that he is going to search he can use the application to get the meaning of that word and then he can search the corresponding string. If he knows that he is going to search with meaning, then he can give the string as it is and can search. This plays a dual part in which depending upon the users choice the search or retrieve process initiates.

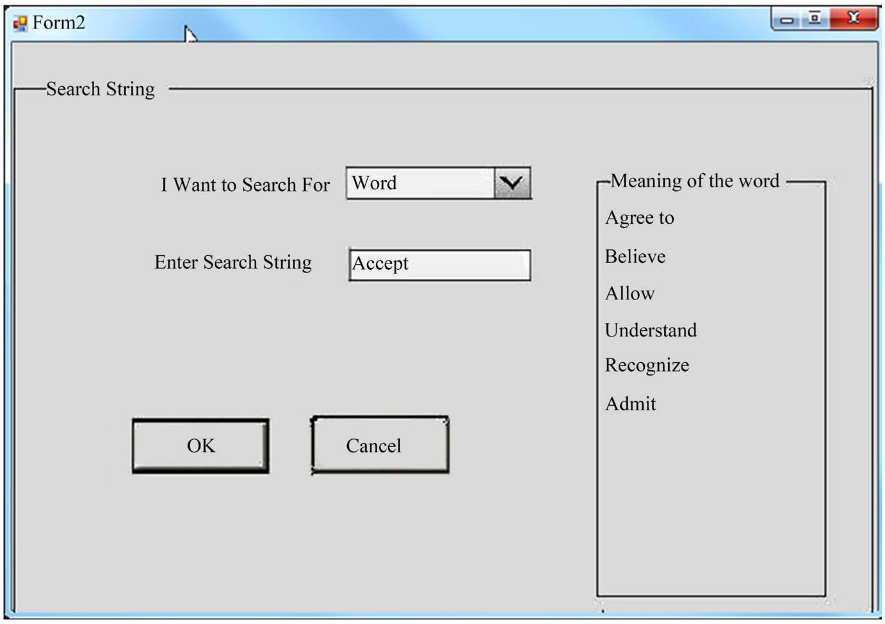
If the user doesn't know the meaning for the word they are going to search, then they can use the inbuilt dictionary within the application. The application fetches the word from the user, identifies the various meanings of the given word. After selecting the required meanings from the database. It displays it to the user. From that, user can know its relevant meaning and can search their required search. Further, when the user gives a document it actually checks each and every word of the document and compares it with the database pre-saved keywords, if it matches with any of those words the type of the document is thus identified as in **Figure 7**. It is a step at which a user initiates to fetch some document. When a user gives a query keyword to search, it receives it, searches in the database for matching documents with desired annotation and displays it.

5.3. Segregation and Annotation of the Document

If the keyword in the document matches with the keywords in the database, that particular paragraph is extracted from the document and if the type of the document is identified by the keyword match—search process, that particular type is taken it is used to annotate the document. If a particular document is having some weather report with those keywords like tropical storm it annotates the document as storm—Tropical storm. The application fetches the word from the user, identifies the various meanings of the given word. After selecting the required meanings from the database, it displays it to the user. From that, user can know its relevant meaning and can search his required search. If the particular document is not based on any particular content type then that entire document is annotated with the common type. For example, If a particular book is concerned with the information of apple company then if that particular book is processed with our software reads the book identifies the keyword like model, version, year, operating system and then identifies that it is related with the apple company. So it annotates the book as apple.

5.4. Document Retrieval

If a particular user gives some queries or keywords to search in order to display a document, it searches with the



Form2

Search String

I Want to Search For: Word

Enter Search String: Accept

Meaning of the word

Agree to

Believe

Allow

Understand

Recognize

Admit

Figure 7. Semantic analysis of the keyword.

relevant type if it is identified or else as a whole it is processed. The user has to give the desired word for search. **Figure 8** shows the process of obtaining the required document based on search.

After that, the search process uses an information extraction algorithm. It retrieves the required documents as either in T [Good] or T [Best]. In general, every day new words may arise. So, the database has to be frequently updated with the words and their meanings.

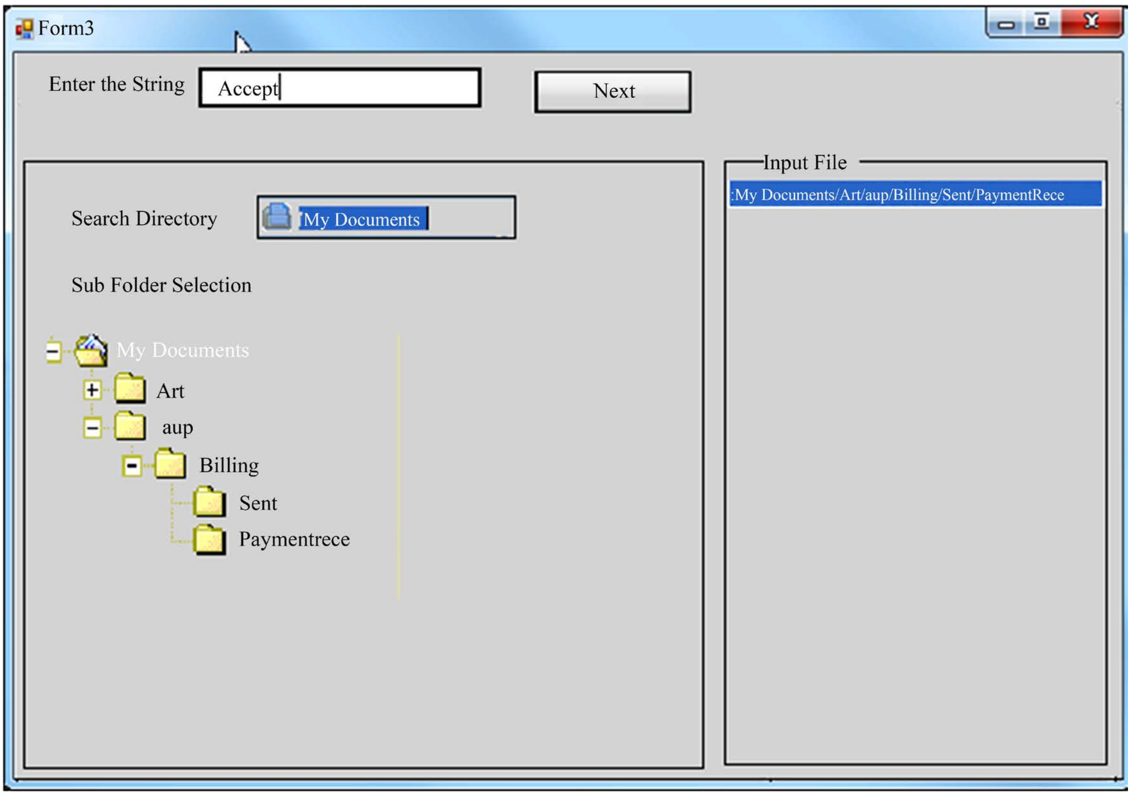
6. Performance Analysis

An option for the user's interest is provided. If a person is interested to search for a meaning, they can give the search string directly. If the choice is word and if they don't know the meaning, they can use the inbuilt search engine to first get the meaning and then they can search. In the search process too, they can limit the search by using the directory and sub folder selection. **Figure 9** illustrates the performance analysis of the document based on the segregation using adaptive progression technique.

In addition, it is going to retrieve the documents alone. So, Unwanted time waste in searching the entire system for the particular word with images, video files and etc are avoided.

7. Conclusion and Future Work

In general, a search processed by an operating system doesn't give importance for the word given. It instead, gives importance for the folders name or files name with which it is stored. It retrieves all the documents, files like images, power point presentations and so on. But, when we use this system to find the documents that exactly matches a word, it retrieves the exact documents that match with the keyword alone. So, unwanted files processing is avoided and it saves the time. The efficiency is more and only the exact documents are retrieved. Time to search the entire hard drive to unrelated documents is time consuming. So, we retrieve the documents alone by using this IE algorithm. It helps us to improve the accessibility for a document. Unwanted time waste in searching all the files will be eliminated. So, it gives the best result. Even if the meaning for a particular word is not known, we can use this application to retrieve the relevant meanings. For such users it will be very useful.



Enter the String: Accept

Search Directory: My Documents

Input File: :My Documents/Art/aup/Billing/Sent/PaymentRece

Figure 8. Obtaining the required document based on search.

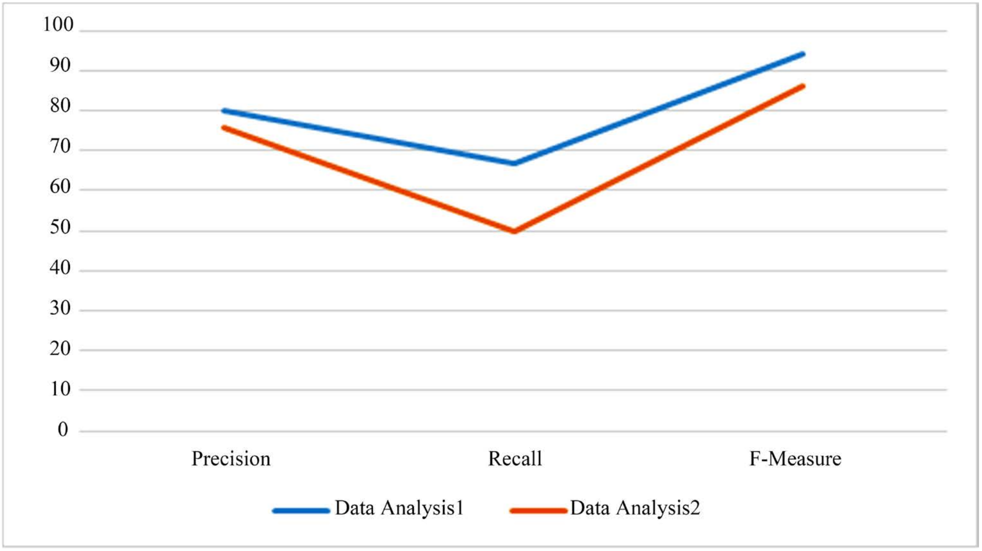


Figure 9. Performance analysis of document.

As a future work, classifying documents as belonging to a particular domain by using the common words that we can use to have in a document can be performed. The domain classification is quite a tedious process because of the up gradation in technology on daily basis. It requires updating of a database on daily basis.

References

[1] (2011) Google. Google Base. http://www.google.com/base

[2] Jeffery, S.R., Franklin, M.J. and Halevy, A.Y. (2008) Pay-as-You-Go User Feedback for Data Space Systems. *SIGMOD*'08 *Proceedings of the* 2008 *ACM SIGMOD International Conference on Management of Data*, 847-860. http://dx.doi.org/10.1145/1376616.1376701

[3] Jain, A. and Ipeirotis, P.G. (2009) A Quality-Aware Optimizer for Information Extraction. *ACM Transactions on Database Systems*, **34**, Article 5.

[4] J.M. Ponte and W.B. Croft (1998) A Language Modeling Approach to Information Retrieval. *Proceedings of the* 21*st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR*'98), ACM, New York, 275-281. http://dx.doi.org/10.1145/290941.291008

[5] Chang, K.C.-C. and Hwang, S.-W. (2002) Minimal Probing: Supporting Expensive Predicates for Top-K Queries. *Proc. ACM SIGMOD International Conference on Management Data*, Madison, Wisconsin, 4-6 June 2002, 12 p. http://dx.doi.org/10.1145/564691.564731

[6] Heymann, P., Ramage, D. and Garcia-Molina, H. (2008) Social Tag Prediction. *Proceedings of the* 31*st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR*'08), ACM, New York, 531-538. http://dx.doi.org/10.1145/1390334.1390425

[7] Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J., Lee, W.C. and Giles, C.L. (2008) Real-Time Automatic Tag Recommendation. *Proc.* 31*st Ann. Int'l ACM SIGIR. Conf. Research and Development in Information Retrieval (SIGIR*'08), 515-522. http://dx.doi.org/10.1145/1390334.1390423

[8] Etzioni, O., Banko, M., Soderland, S. and Weld, D.S. (2008) Open Information Extraction from the Web. *Communications of the ACM*, **51**, 68-74. http://dx.doi.org/10.1145/1409360.1409378

[9] Doan, A., Ramakrishnan, R., Chen, F., DeRose, P., Lee, Y., McCann, R., Sayyadian, M. and Shen, W. (2006) Community Information Management. *IEEE Data Engineering Bulletin*, **29**, 64-72.

[10] Chu, E., Baid, A., Chai, X., Doan, A. and Naughton, J. (2009) Combining Keyword Search and Forms for Ad Hoc Querying of Databases. *Proceedings of ACM SIGMOD International Conference on Management Data*, 349-360. http://dx.doi.org/10.1145/1559845.1559883

[11] Banerjee, J., Kim, W., Kim, H.J. and Korth, H.F. (1987) Semantics and Implementation of Schema Evolution in Object-Oriented Databases. *Proceedings of ACM SIGMOD International Conference on Management Data*, **16**, 311-322.

[12] Nandi, A. and Jagadish, H.V. (2007) Assisted Querying Using Instant-Response Interfaces. *Proceedings of ACM*

SIGMOD International Conference on Management Data, 472-483. http://dx.doi.org/10.1145/1247480.1247640

[13] Yin, D., Xue, Z., Hong, L. and Davison, B.D. (2010) A Probabilistic Model for Personalized Tag Prediction. *Proceedings of ACM SIGMOD International Conference on Knowledge Discovery Data Mining*, Washington, DC, July 2010, 959-968. http://dx.doi.org/10.1145/1835804.1835925