Cluster Based Web Search Using Support Vector Machine

Rita. S. Shelke ritashelke@gmail.com
Department of Computer Engineering
Bharati Vidyapeeth University, COE
Pune-43, India.

Devendra Singh Thakore deventhakore@yahoo.com
Department of Computer Engineering
Bharati Vidyapeeth University, COE
Pune-43, India

Abstract

Now days, searches for the web pages of a person with a given name constitute a notable fraction of queries to Web search engines. This method exploits a variety of semantic information extracted from web pages.
The rapid growth of the Internet has made the Web a popular place for collecting information. Today, Internet user access billions of web pages online using search engines. Information in the Web comes from many sources, including websites of companies, organizations, communications and personal homepages, etc. Effective representation of Web search results remains an open problem in the Information Retrieval community. For ambiguous queries, a traditional approach is to organize search results into groups (clusters), one for each meaning of the query. These groups are usually constructed according to the topical similarity of the retrieved documents, but it is possible for documents to be totally dissimilar and still correspond to the same meaning of the query. To overcome this problem, the relevant Web pages are often located close to each other in the Web graph of hyperlinks. It presents a graphical approach for entity resolution & complements the traditional methodology with the analysis of the entity-relationship (ER) graph constructed for the dataset being analyzed. It also demonstrates a technique that measures the degree of interconnectedness between various pairs of nodes in the graph. It can significantly improve the quality of entity resolution.

Using Support Vector Machines (SVMs) which are a set of related Supervised learning methods used for classification of load of user queries to the sever machine to different client machines so that system will be stable. Cluster web pages based on their capacities stores whole database on server machine.

Keywords: SVM, Cluster, ER.

1. INTRODUCTION

Web search is difficult because it is hard for users to construct queries that are both sufficiently descriptive and sufficiently discriminating to find just the web pages that are relevant to the user's search goal. Queries are often ambiguous: words and phrases are frequently polysemantic and user search goals are often narrower in scope than the queries used to express them. This ambiguity leads to search result sets containing distinct page groups that meet different user search goals. Often users must refine their search by modifying the query to filter out the irrelevant results. Users must understand the result set to refine queries effectively; but this is time consuming, if the result set is unorganized. Web page clustering is one approach for assisting users to both comprehend the result set and to refine the query. Web page clustering identifies semantically meaningful groups of web pages and presents these to the user as clusters. The clusters provide an overview of the contents of the result set and when a cluster is selected the result set is refined to just the relevant pages in that cluster. After clustering

whatever the load of queries on a single machine which is treated as server will get distributed over the network using Support Vector Machine, which act as load classifier or distributor. Depending upon the capacities of machines each can handle the specific load & return unstable when it exceed the limit. It is useful to determine whether the machine is stable or not.

1.1 Present Theories & Practices

There has been a large body of work on web search, unambiguation, entity resolution. Here we review some of the main work, but the review is not exhaustive. In Web people search application [1] the main techniques used are unambiguation & entity resolution. The authors have overviewed several existing entity resolution approaches, pointing out that they rely primarily on analyzing object features for making their co reference decisions. As compared it with existing unambiguation works. The Authors have developed a novel algorithm for unambiguating among people that have the same name which is based on extracting "significant" entities such as the names of other persons, organizations, and locations on each web page, forming relationships between the person associated with the web page and the entities extracted, and then analyzing the relationships along with features such as TF/IDF, as well as other useful content including hyperlink information to disambiguate the pages. Then design a cluster-based people search approach based on the disambiguation algorithm. In Disambiguation Algorithm for People Search on the Web [2], the authors have concentrated on disambiguation algorithm which exist for a variety of data management applications. The proposed disambiguation algorithm is based on analyzing two types of information. First, it analyzes object features, like many other techniques. Second, (most important) it also analyzes the Entity-Relationship Graph (ER graph) for the dataset. The idea behind analyzing features of objects u, v is based on the assumption that similarity of features of two objects defines certain affinity/attraction between those objects f(u, v). If this attraction f(u, v) is sufficiently large, then the objects are likely to be the same (co-refer). The intuition behind analyzing paths in the ER graph is similar. The assumption is that each path/connection/link p between two objects u, v can serve as evidence that they co-refer. So if the combined evidence, stored in all the u-v paths, is sufficiently large, the objects are likely to be the same. Formally, the attraction between two nodes u and v via paths is measured using the connection strength measure c(u, v) which is defined as the sum of attractions contributed by each path:

$$c(u,v) = \sum_{p \in P_{uv}} w_p.$$

Here P_{uv} denotes the set of all simple paths between u and v (of limited length), and c(p) is the contribution of path p.

1.2 The Limitations Of Web Search

With an enormous growth of the Internet it has become very difficult for the users to find relevant documents. In response to the user's query, currently available search engines return a ranked list of documents along with their partial content. If the query is general, it is extremely difficult to identify the specific document which the user is interested in. The users are forced to sift through a long list of off-topic documents. Moreover, internal relationships among the documents in the search result are rarely presented and are left for the user. Standard information retrieval systems rely on two orthogonal paradigms: the textual similarity with the query (e.g., tf-idf-based cosine similarity) on one hand and a query independent measure of each web page's importance (e.g., link authority ranking) on the other hand. However, these systems generally lack user modeling and thus are far from being optimal i.e Different users may submit exactly the same query even though they have different intentions. The most famous examples of such ambiguous queries include bass (fish or instrument), java (programming language, island or coffee), jaguar (animal, car or Apple software) and IR application (Infrared application or Information Retrieval application) For example assume we have two users, one of whom is a computer science student, and the other is geography student. Figure 1 represents the top-6 results returned by Google when the query "java map" is submitted. The result set spans two categories, namely the java map collection classes and maps for the Indonesian island java. Generally speaking, the computer science student would be most likely interested in the java map collection classes,

where as the geography student would be interested in locating maps for the Indonesian island java.

FIGURE 1: Overview of query processing

1.3 Cluster _Based Web Search

Clustering of Web search results has been in the focus of IR community since the yearly days of the Web. There are two reasons for clustering of search results. The first is that the IR research community has long recognized the validity of the clustering approach in top ranked documents; I. e. similar documents tend to be relevant to the same request. A second reason is that the ranked list is usually too large and contains many documents that are irrelevant to the particular meaning of the query the user had in mind. Thus it would be beneficial to group search results by various meanings of the query. The attempts have made the clustering of search results for many web users. However, it is still not accurate enough to attract an average user. The main drawback of many Web page clustering methods is that they take into account only the topical similarity between the documents in the ranked list. Topical similarity metrics between Web pages would not help solving the clustering problem in at least two cases: (a) when there is not enough contextual information on a page (b) When Websites are contextually different but actually refer to the meaning of the query.

The limitations are as follows:

a. Users must sift through a long list of documents, some of which are irrelevant

b. The reason a document was included in the results is not explicit

c. The relation of a document to the query is not explicit

d. No explicit information is provided about the relationships between documents on the list

e. All documents on the list must be sorted even though some of them may not relate to each other and are thus not comparable.

The solution is that for each such web page, the search-engine could determine which real entity the page refers to. This information can be used to provide a capability of clustered search, where instead of a list of web pages of (possibly) multiple entities with the same name, the results are clustered by associating each cluster to a real entity. The clusters can be returned in a ranked order determined by aggregating the rank of the web pages that constitute the cluster. With each cluster, we also provide a summary description that is representative of the real entity associated with that cluster.

2. QUERY PROCESSING

To overcome these limitations, the goal is to group all the entity descriptions that refer to the same real world entities. A user submits a query to the middle ware via a specialized Web-based interface. The middle ware queries a search engine with this query via the search engine API and retrieves a fixed number (top K) of relevant web pages. The result is a set of clusters of these pages with the aim being to cluster web pages based on association to real entity. Each resulting cluster is then processed. A set of keywords that represent the web pages within a cluster is computed for each cluster. The goal is that the user should be able to find the person of interest by looking at the sketch. The proposed work has been divided into four modules which are 1.Web pages retrieval for the query 2.Preprocessing of web pages 3. Clustering & its Processing 4. Graph Creation.

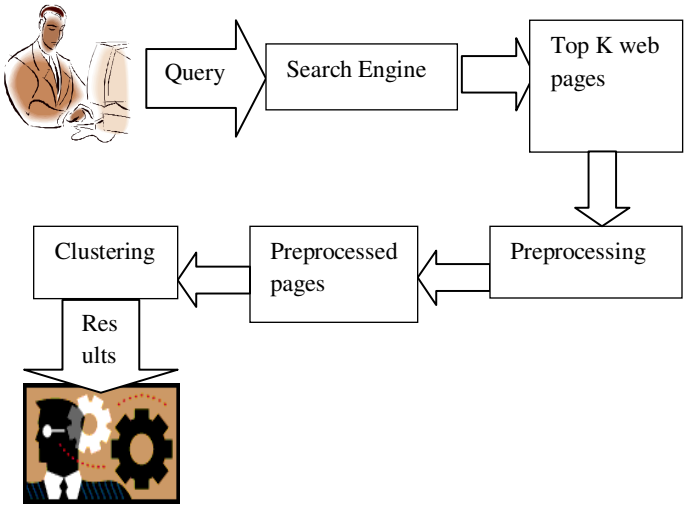

FIGURE 2: Overview of query processing

Web search applications can be implemented in two different settings.

1. Server-side setting
2. Middle ware setting

In server-side setting, the disambiguation mechanism is integrated into the search engine directly. On other hand in a middle ware approach, build entity search capabilities on top of an existing

search-engine such as Google by “wrapping” the original engine. The middle ware would take a user query, use the search engine API to retrieve top K web pages most relevant to the user query, and then cluster those web pages based on their associations to real people. The middle-ware approach is more common, as it is difficult to conduct realistic testing of the server-side approach due to the lack of direct access to the search engine internal data. The architecture is a pipeline that receives the input query, obtains search results from a search engine, filters the results applying a clustering algorithm and then gets the clusters. The steps of overall approach are illustrated in Figure 3.

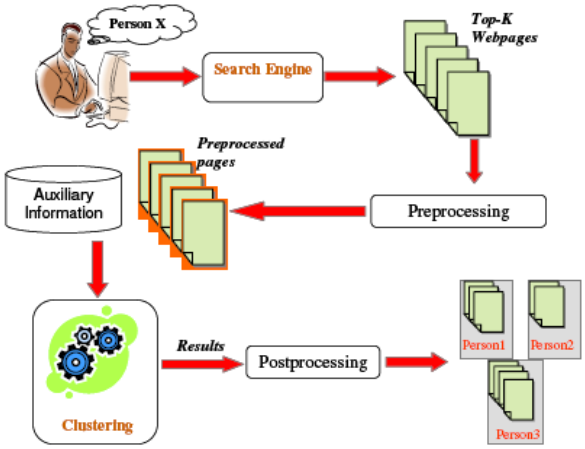

FIGURE 3: Overview of processing steps

2.1 Design Of Web Pages Retrieval For Query

Web Pages retrieval for query can be implemented in many ways. There are many algorithms to process Top-k retrieval, for example: Fagins Threshold Algorithm (TA), No Random Access Algorithm (NRA) and Combined Algorithm (CA). All these threshold algorithms work on inverted indices for query terms. Assuming the vector space model, the way to fetch the top-k documents would be to compute the textual similarity of all the documents present in the corpus with the query vector, order them according to this similarity score and then fetch the top-k documents from this ordered list. However taking into consideration the huge size of the web corpus, this process becomes very unfeasible. The *HttpServlet* component seeks to fill this void by providing an efficient, up-to-date, and feature-rich package implementing the client side of the most recent HTTP standards and recommendations. The features are standards based, pure Java, implementation of HTTP versions 1.0 and 1.1. It is the full implementation of all HTTP methods. The figure 4, shows the process of retrieving the top pages from the search engine.

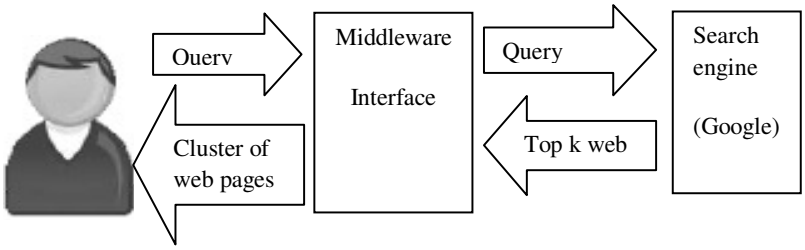

FIGURE 4: Web pages retrieval

2.2 Design of Preprocessing Of Web Pages

After retrieving the top pages related to the query, the pages are processed by using IR techniques. There are various algorithms which are simply a set of instructions, usually mathematical, used to calculate a certain parameter and perform some type of data processing.

The job is to generate a set of highly relevant documents for any search query, using the available parameters on the web. The task is challenging because the available parameters usable by the algorithm are not necessarily the same as the ones web users see when deciding if a webpage is relevant to their search. The figure 5 shows the preprocessing of the web pages which include the two processes named as stemming & stop word removal.

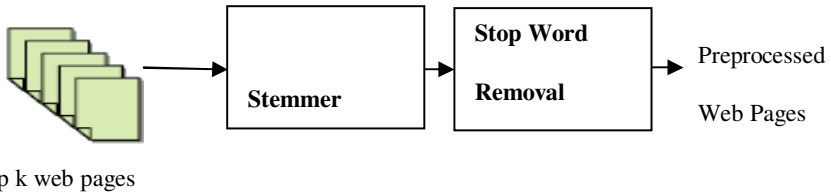

FIGURE 5: Processing of Web Pages

2.2.1 Stemming

Stemming algorithms are used to transform the words in texts into their grammatical root form, and are mainly used to improve the Information Retrieval System's efficiency. To stem a word is to reduce it to a more general form, possibly its root. For example, stemming the term interesting may produce the term interest. Though the stem of a word might not be its root, we want all words that have the same stem to have the same root. The effect of stemming on searches of English document collections has been tested extensively. Several algorithms exist with different techniques. The most widely used is the Porter Stemming algorithm. In some contexts, stemmers such as the Porter stemmer improve precision/recall scores [3]. The stemmer operations are classified into rules where each of these rules deals with a specific suffix and having certain condition(s) to satisfy. A given word's suffix is checked against each rule in a sequential manner until it matches one, and consequently the conditions in the rule are tested on the stem that may result in a suffix removal or modification. Using $(VC)^m$ to denote VC repeated m times, this may again be written as $[C](VC)^m[V]$.m will be called the measure of any word or word part when represented in this form. The case m = 0 covers the null word. The algorithm now follows:

Step 1a

Rules	Illustrations
SSES -> SS	caresses ->caress
IES -> I	ponies -> poni, Ties ->ti
SS -> SS	caress -> caress
S ->	cats -> cat

Step 1b

Rules	Illustrations
(m>0) EED -> EE	feed -> feed, agreed ->agree
(*v*) ED ->	plastered -> plaster, bled -> bled
(*v*) ING ->	motoring -> motor, sing -> sing

If the second or third of the rules in Step 1b is successful, the following is done:

AT -> ATE	conflat(ed) -> conflate
BL -> BLE	troubl(ed) -> trouble
IZ -> IZE	siz(ed) -> size

(*d and not (*L or *S or *Z))

-> single letter

hopp(ing) -> hop, tann(ed) -> tan

(m=1 and *o) -> E fail(ing) -> fail, fil(ing) -> file

The rule to map to a single letter causes the removal of one of the double letter pair. The -E is put back on -AT, -BL and -IZ, so that the suffixes -ATE, -BLE and -IZE can be recognized later. This E may be removed in step 4.

Step 1c

Rule	Illustrations
(*v*) Y -> I	happy -> happi, sky -> sky

Step 1 deal with plurals and past participles. The subsequent steps are much more straightforward.

Step 2

Rules	Illustrations
(m>0) ATIONAL -> ATE	relational -> relate
(m>0) TIONAL -> TION	conditional -> condition, rational->rational
(m>0) ENCI -> ENCE	valenci -> valence
(m>0) ANCI -> ANCE	hesitanci -> hesitance
(m>0) IZER -> IZE	digitizer -> digitize
(m>0) ABLI -> ABLE	conformabli -> conformable
(m>0) ALLI -> AL	radicalli -> radical
(m>0) ENTLI -> ENT	differentli -> different
(m>0) ELI -> E	vileli - > vile
(m>0) OUSLI -> OUS	analogousli -> analogous
(m>0) IZATION -> IZE	vietnamization -> vietnamize
(m>0) ATION -> ATE	predication -> predicate
(m>0) ATOR -> ATE	operator -> operate
(m>0) ALISM -> AL	feudalism -> feudal
(m>0) IVENESS -> IVE	decisiveness -> decisive
(m>0) FULNESS -> FUL	hopefulness -> hopeful
(m>0) OUSNESS -> OUS	callousness -> callous

(m>0) ALITI -> AL	formaliti -> formal
(m>0) IVITI -> IVE	sensitiviti -> sensitive
(m>0) BILITI -> BLE	sensibiliti -> sensible

Step 3

Rules	Illustrations
(m>0) ICATE -> IC	triplicate -> triplic
(m>0) ATIVE ->	formative -> form
(m>0) ALIZE -> AL	formalize -> formal
(m>0) ICITI -> IC	electriciti -> electric
(m>0) ICAL -> IC	electrical -> electric
(m>0) FUL ->	hopeful -> hope
(m>0) NESS ->	goodness -> good

Step 4

Rules	Illustrations
(m>1) AL ->	revival -> reviv
(m>1) ANCE ->	allowance -> allow
(m>1) ENCE ->	inference -> infer
(m>1) ER ->	airliner -> airlin
(m>1) IC ->	gyroscopic -> gyroscop
(m>1) ABLE ->	adjustable -> adjust
(m>1) IBLE ->	defensible -> defens
(m>1) ANT ->	irritant -> irrit
(m>1) EMENT ->	replacement -> replac
(m>1) MENT ->	adjustment -> adjust
(m>1) ENT ->	dependent -> depend
(m>1 and (*S or *T)) ION ->	adoption -> adopt
(m>1) OU ->	homologou -> homolog
(m>1) ISM ->	communism -> commun
(m>1) ATE ->	activate -> activ

(m>1) ITI -> angulariti -> angular

(m>1) OUS -> homologous -> homolog

(m>1) IVE -> effective -> effect

(m>1) IZE -> bowdlerize -> bowdler

The suffixes are now removed. All that remains is a little tidying up.

Step 5a

Rules	Illustrations
(m>1) E ->	probate -> probat, rate->rate
(m=1 and not *o) E ->	cease -> ceas

Step 5b

Rules	Illustrations
(m > 1 and *d and *L) -> single letter	control->control

The algorithm is careful not to remove a suffix when the stem is too short, the length of the stem being given by its measure, m. It was merely observed that m could be used quite effectively to help decide whether or not it was wise to take off a suffix.

2.2.2 Elimination of Stop Words

After stemming it is necessary to remove unwanted words. There are 400 to 500 types of stop words such as “of”, “and”, “the,” etc., that provide no useful information about the document’s topic. Stop-word removal is the process of removing these words. Stop-words account for about 20% of all words in a typical document[4]. These techniques greatly reduce the size of the search engine’s index. Stemming alone can reduce the size of an index by nearly 40%. To compare a webpage with another webpage, all unnecessary content must be removed and the text put into an array. The Sam Allen has proposed the stop word dictionary on Dot Net Perls but this is not sufficient for some of the applications. For this reason, static dictionary is modified.

2.3 Design of Clustering & Its Processing

When designing a Cluster Based Web Search, special attention must be paid to ensuring that both content and description (labels) of the resulting groups are meaningful to humans. As stated, “a good cluster—or document grouping—is one, which possesses a good, readable description”. There are various algorithms such as K means, K-medoid but this algorithm require as input the number of clusters. A Correlation Clustering (CC) algorithm is employed which utilizes supervised learning. The key feature of Correlation Clustering (CC) algorithm is that it generates the number of clusters based on the labeling itself & not necessary to give it as input but it is best suitable when query is person names. For general query, the algorithms are Query Directed Web Page Clustering (QDC), Suffix Tree Clustering (STC), Lingo, and Semantic Online Hierarchical Clustering (SHOC).The focus is made on Lingo because the QDC considers only the single words. The STC tends to remove longer high quality phrases, leaving only less informative & shorter ones. So, if a document does not include any of the extracted phrases it will not be included in results although it may still be relevant. To overcome the STC's low quality phrases problem, in SHOC introduce two novel concepts: complete phrases and a continuous cluster definition. The drawback of SHOC is that it provides vague threshold value which is used to describe the resulting cluster. Also in many cases, it produces unintuitive continuous clusters. The majority of open text clustering algorithms follows a scheme where cluster content discovery

is performed first, and then, based on the content, the labels are determined. But very often intricate measures of similarity among documents do not correspond well with plain human understanding of what a cluster's "glue" element has been. To avoid such problems Lingo reverses this process—first attempt to ensure that we can create a human-perceivable cluster label and only then assign documents to it. Specifically, extract frequent phrases from the input documents, hoping they are the most informative source of human-readable topic descriptions. Next, by performing reduction of the original term-document matrix using Singular Value Decomposition (SVD), try to discover any existing latent structure of diverse topics in the search result. Finally, match group descriptions with the extracted topics and assign relevant documents to them.

LINGO – Main phase's pseudo-code

```
Phase 1: Preprocessing
 for each document
{
 apply stemming;
 mark stop words;
 }
Phase 2: Frequent Phrase Extraction
   discover frequent terms and phrases;
Phase 3: Cluster label induction
     use LSI to discover abstract concepts;
   for each abstract concept
   {
   find best-matching phrase;
   }
    prune similar cluster labels;
Phase 4: Cluster content discovery
   for each cluster label
  {
    use VSM to determine the cluster contents;
  }
Phase 5: Final cluster formation
     calculate cluster scores;
      apply cluster merging;
```

3. FREQUENT PHRASE EXTRACTIONS

The frequent phrases are defined as recurring ordered sequences of terms appearing in the input documents. Intuitively, when writing about something, we usually repeat the subject-related keywords to keep a reader's attention. Obviously, in a good writing style it is common to use synonymy and pronouns and thus avoid annoying repetition. The Lingo can partially overcome the former by using the SVD-decomposed term document matrix to identify abstract concepts—single subjects or groups of related subjects that are cognitively different from other abstract concepts. To be a candidate for a cluster label, a frequent phrase or a single term must:

1. Appear in the input documents at least certain number of times (term frequency threshold),

2. Not cross sentence boundaries,

3. Be a complete phrase (see definition below),

4. Not begin nor end with a stop word.

A complete phrase is a complete substring of the collated text of the input documents, defined in the following way: Let T be a sequence of elements (t_1, t_2, t_3 . . . t_n). S is a complete substring of T when S occurs in k distinct positions p_1, p_2, p_3 . . . p_k in T and Ǝi, j Є 1 . . . k : $t_{pi-1} \neq t_{pj-1}$ (left-completeness) and Ǝi, j Є 1 . . . k : $t_{pi+|S|} \neq t_{pj+|S|}$ (right-completeness). In other words, a complete phrase cannot be "extended" by adding preceding or trailing elements, because at least one of these elements is different from the rest. An efficient algorithm for discovering complete phrases was proposed in [5], although it contained one mistake that caused the frequency of some phrases to be miscalculated. It does not affect further discussion of Lingo because any algorithm capable of discovering frequent phrases could be used at this stage. Figure 6 presents the whole phrase extraction phases.

LINGO – phrase extraction phase pseudo-code

```
Phase 2: Frequent phrases extraction

        Conversion of the representation

  for each document

{   convert the document from the character-based to

    the word-based representation;

}

        Document concatenation

concatenate all documents;

create an inverted version of the concatenated documents;

        Complete phrase discovery

discover right-complete phrases;

discover left-complete phrases;

sort the left-complete phrases alphabetically;

combine the left- and right-complete phrases into a set of complete phrases;
```

Final selection

for further processing choose the terms and phrases whose

frequency exceed the Term Frequency Threshold;

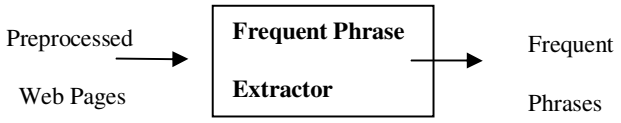

FIGURE 6: Frequent phrase extraction

3.1 Cluster Label Induction

Once frequent phrases (and single frequent terms) that exceed term frequency thresholds are known, they are used for cluster label induction. There are three steps to this: term-document matrix building, abstract concept discovery, phrase matching and label pruning.

The term-document matrix is constructed out of single terms that exceed a predefined term frequency threshold. Weight of each term is calculated using the standard term frequency, inverse document frequency (tf-idf) formula [7], terms appearing in document titles are additionally scaled by a constant factor. In abstract concept discovery, Singular Value Decomposition method is applied to the term-document matrix to find its orthogonal basis. The vectors of this basis (SVD's U matrix) supposedly represent the abstract concepts appearing in the input documents. We estimate the value of k by selecting the Frobenius norms of the term-document matrix A and its k-rank approximation A_k. Let threshold q be a percentage-expressed value that determines to what extent the k-rank approximation should retain the original information in matrix A. We hence define k as the minimum value that satisfies the following condition: $||A_k||F / ||A||F \geq q$, where $||X||F$ symbol denotes the Frobenius norm of matrix X. Clearly, the larger the value of q the more cluster candidates will be induced. The choice of the optimal value for this parameter ultimately depends on the users' preferences. Therefore make it one of Lingo's control thresholds—Candidate Label Threshold.

Phrase matching and label pruning step, where group descriptions are discovered, relies on an important observation that both abstract concepts and frequent phrases are expressed in the same vector space—the column space of the original term-document matrix A. Thus, the classic cosine distance can be used to calculate how "close" a phrase or a single term is to an abstract concept. Let us denote by P a matrix of size $t \times (p+t)$ where t is the number of frequent terms and p is the number of frequent phrases. P can be easily built by treating phrases and keywords as pseudo-documents and using one of the term weighting schemes. Having the P matrix and the i^{th} column vector of the SVD's U matrix, a vector m_i of cosines of the angles between the i^{th} abstract concept vector and the phrase vectors can be calculated: $_i = U_i^T P$. The phrase that corresponds to the maximum component of the m_i vector should be selected as the human-readable description of i^{th} abstract concept. Additionally, the value of the cosine becomes the score of the cluster label candidate. A similar process for a single abstract concept can be extended to the entire U_k matrix—a single matrix multiplication $M = U_k^T P$ yields the result for all pairs of abstract concepts and frequent phrases. On one hand we want to generalize information from separate documents, but on the other we want to make it as narrow as possible at the cluster description level. Thus, the final step of label induction is to prune overlapping label descriptions. Let V be a vector of cluster label candidates and their scores. We create another term-document matrix Z, where cluster label candidates serve as documents. After column length normalization calculates $Z^T Z$, which yields a matrix of similarities between cluster labels. For each row we then pick columns that exceed the Label Similarity Threshold and discard all but one cluster label candidate with the maximum score.

LINGO – cluster label induction phase pseudo-code

Phase 3: Cluster label induction

Term-document matrix building

build the term-document matrix A for the input snippet collection.

as index terms use the non-stop words that exceed the predefined

term frequency threshold. use the tf-idf weighting scheme;

Abstract concept discovery

perform the Singular Value Decomposition of the term-document

matrix to obtain U, S and V matrices;

based on the value of the q parameter and using the S matrix -

calculate the desired number k of abstract concepts;

use the first k columns of the U matrix to form the U_k matrix;

Phrase matching

using the tf-idf term weighting create the phrase matrix P;

for each column of the U_k matrix

{

multiplicate the column by the P matrix;

find the largest value in the resulting vector to determine

the best matching phrase;

}

Candidate label pruning

calculate similarities between all pairs of candidate labels;

form groups of labels that exceed a predefined similarity threshold;

for each group of similar labels

{

select one label with the highest score;

}

3.2. Cluster Content Discovery

In the cluster content discovery phase, the classic Vector Space Model (VSM) is used to assign the input documents to the cluster labels induced in the previous phase. In a way, re-query the input document set with all induced cluster labels. The assignment process resembles document retrieval based on the VSM model. Let us define matrix Q, in which each cluster label is represented as a column vector. Let $C = Q^T A$, where A is the original term-document matrix for input documents. This way, element c_{ij} of the C matrix indicates the strength of membership of the j^{th} document to the i^{th} cluster. A document is added to a cluster if c_{ij} exceeds the Snippet Assignment Threshold, yet another control parameter of the algorithm. Documents not assigned to any cluster end up in an artificial cluster called others.

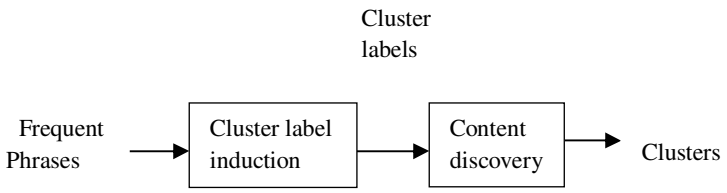

FIGURE 7: Cluster Formation

3.3 Final Cluster Formation

Finally, clusters are sorted for display based on their score, calculated using the following simple formula: C_{score} = label score × ||C||, where ||C|| is the number of documents assigned to cluster C. The scoring function, although simple, prefers well-described and relatively large groups over smaller, possibly noisy ones. For the time being, no cluster merging strategy or hierarchy induction is used for Lingo.

3.4 Design of Graph Creation

It is a graphical approach, as it visualizes the dataset as the standard entity-relationship graph. There are other graphical disambiguation approaches, which visualize different graphs: Web Graph, Co-reference dependence graph, Entity-relationship graph (ER graph). Existing techniques are frequently based on probabilistic methodologies, application rely primarily on the mathematical apparatus from the area of Operation Research. The suitable visualization is the ER graph. By using JGraph class objects and their relations are displayed. A JGraph object doesn't actually contain the data; it simply provides a view of the data. Like any non-trivial Swing component, the graph gets data by querying its data model. Table1 shows the excepted results/cluster names for various categories of queries.

Type of query	Query	Clusters
Ambiguous	Mouse	Computer mouse, Magic mouse, Cursor, gene, House Mouse, Mickey Mouse
General	Music	New music, music news, pop music, songs, Albums, Games
Compound Query	Travel to shimla	Tourism Travels, Travel Guide, shimla Tour Packages, Map, Hotels, Resorts
People Name	Pratibha Patil	Female President, President of India, Governor of Rajasthan, Photos, Videos, Visit

TABLE 1: Clusters

4. TOOLS USED

For many reasons we have decided to implement the system in Java (JDK 1.6). Firstly, with its emphasis on Object Oriented Programming (OOP), Java enforces the much needed good software practices such as the use of interfaces and proper code organization. Finally, Java comes with a package of optimized ready-to-use data structures (such as hash tables or lists) that make programming faster and less error-prone.

4.1 Public Interface HttpServletRequest

The HttpServlet is an abstract class that simplifies writing HTTP servlets. It extends the GenericServlet base class and provides a framework for handling the HTTP protocol. Because it is an abstract class, servlet writers must subclass it and override at least one method. The methods normally overridden are doGet, doPost, doPut.The doGet Performs the HTTP GET operation; the default implementation reports an HTTP BAD_REQUEST error. Overriding this method to support the GET operation also automatically supports the HEAD operation. (HEAD is a GET that returns nobody in the response; it just returns the request HEADer fields.) Servlet writers who override this method should read any data from the request, set entity headers in the response, access the writer or output stream, and, finally, write any response data. The headers that are set should include content type, and encoding. If a writer is to be used to write response data, the content type must be set before the writer is accessed. In general, the servlet implementer must write the headers before the response data because the headers can be flushed at any time after the data starts to be written. Setting content length allows the servlet to take advantage of HTTP "connection keep alive". If content length cannot be set in advance, the performance penalties associated with not using keep alive will sometimes be avoided if the response entity fits in an internal buffer. Entity data written for a HEAD request is ignored. Servlet writers can, as a simple performance optimization, omit writing response data for HEAD methods. If no response data is to be written, then the content length field must be set explicitly. The GET operation is expected to be safe: without any side effects for which users might be held responsible. For example, most form queries have no side effects. Requests intended to change stored data should use some other HTTP method. The GET operation is also expected to be idempotent: it can safely be repeated. This is not quite the same as being safe, but in some common examples the requirements have the same result. For example, repeating queries is both safe and idempotent (unless payment is required!), but buying something or modifying data is neither safe nor idempotent.

protected void doGet (HttpServletRequest req,

HttpServletResponse resp) throws ServletException, IOException

Parameters:

req - HttpServletRequest that encapsulates the request to the servlet

resp - HttpServletResponse that encapsulates the response from the servlet

Throws: IOException

if detected when handling the request

Throws: ServletException

if the request could not be handled

4.2 JGraph

JGraph is a mature, feature-rich open source graph visualization library written in Java.JGraph is written to be a fully Swing compatible component, both visually and in its design architecture. JGraph can be run on any system supporting Java version 1.4 or later. JGraph provides a range of graph drawing functionality for web applications. JGraph has a simple, yet powerful API enabling you to visualize, interact with, automatically layout and perform analysis of graphs. JGraph, through it's programming API, provides the means to configure how the graph or network is displayed and the means to associate a context or metadata with those displayed elements. JGraph visualization is based on the mathematical theory of networks, graph theory. The main package is JGraph itself which comprises the basic JGraph swing component:

Java Package Name	Functionality
org.jgraph	Basic JGraph class
org.jgraph.event	Graph Event Models
org.jgraph.graph	Graph Structure and nodes
org.jgraph.plaf	Graph UI delegate component
org.jgraph.util	General purpose utilities
com.jgraph.algebra	Graph Analysis Routines
com.jgraph.layout	JGraph Facade and utilities
com.jgraph.layout.organic	Force directed layouts
com.jgraph.layout.tree	Tree layouts
com.jgraph.layout.routing	Edge routing algorithms
com.jgraph.layout.hierarchical	Hierarchical layouts

TABLE 2: JGraph Package

5. ADVANTAGES

1. Readable cluster labels: - The use of phrases in the process of cluster label induction guarantees that group descriptions can be easily understood by the users. As argued in [8], frequent phrases significantly increase the overall quality of clustering, not only of the phrase-based algorithms (such as Suffix Tree Clustering) but also of other approaches such as k-means. Similar effects can be observed also in Cluster Based Web Search.
2. Diverse cluster labels: - Apart from the general abstract concepts related to fairly large groups of documents, Latent Semantic Indexing discovers narrower, more-specific ones. In this way meaningful clusters can be created whose labels are not necessarily the highest-frequency phrases. Additionally, the orthogonality of the SVD-derived abstract concept vectors makes the diversity among cluster labels even wider.

3.Overlapping clusters:- Placing the same document in a larger number of clusters increases the chances that, viewing only selected groups, the user is be able to identify all relevant documents. Moreover, some snippets may be related to two or more topics and thus should be placed in all respective clusters.

4. Modular design: - As all the phases of system are easily separable. Thus, it is possible to provide alternative implementations of some of them, improving the quality or time-efficiency of the algorithm as a whole.

6. RESULTS

The system was implemented using Netbean 6.5.1 as development tool & Jdk 1.6 development Platform .Also it was tested for variety of queries under following four categories and the results obtained where satisfactory.

6.1 System Interface

This module gives the facilities for specifying the various queries to the middleware. The front end developed so far is as follows. The Figure 8 shows user interface, by using that the user enters the query to the middleware. Along with the query, user can also select the number of results (50/100/150/200) to be fetched from source. In Figure 8, query entered is “mouse” & result selected is 100.

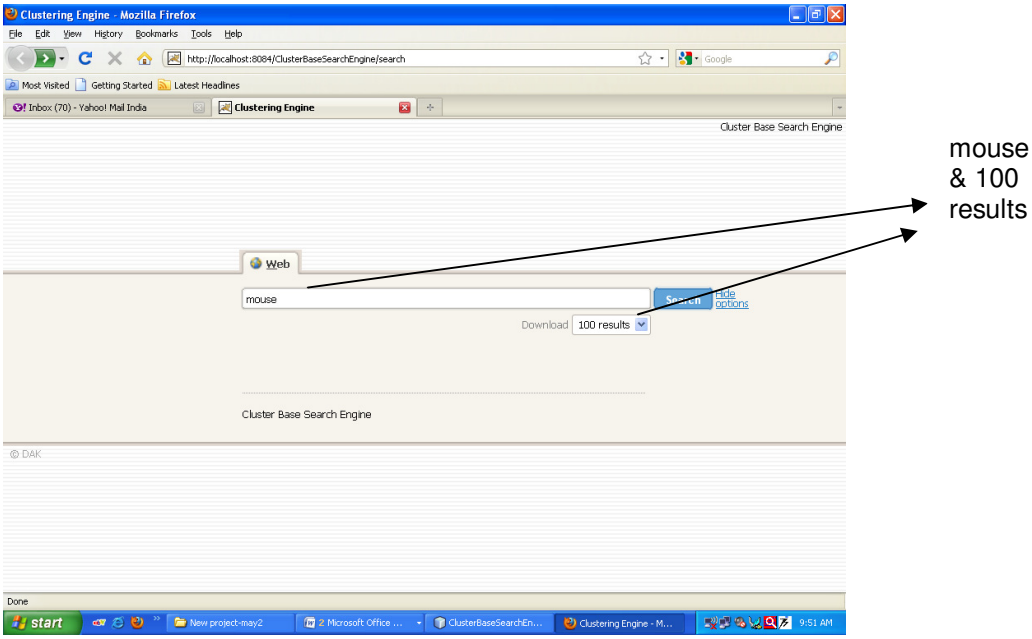

FIGURE 8: System Interface

6.2 Web Pages Retrieval for the Query

The user issues a query to the system (middleware) sends a query to a search engine, such as Google, and retrieves the top-K returned web pages. This is a standard step performed by most of the current systems. The figure 9 shows that the 200 results were fetched from the source Google for query “mouse”.

Input: Query “mouse” & k=50/100/150/200 pages Output: Web pages of Query “mouse”

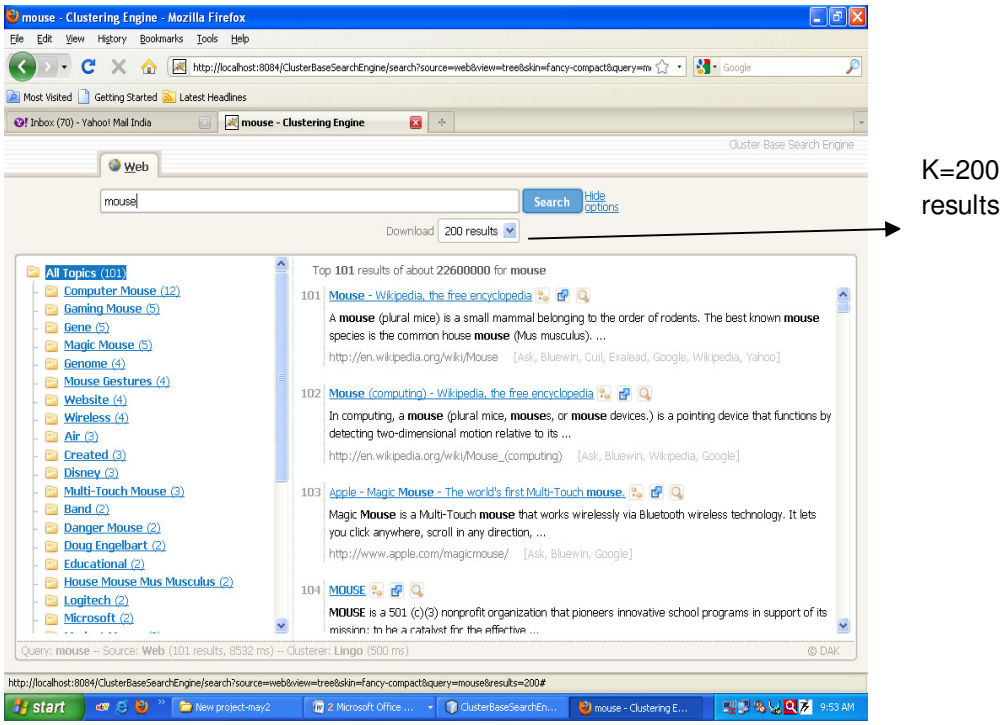

FIGURE 9: Clustering results for a ambiguous query “mouse” & k=200 result

6.3 Preprocessing Of Web Pages

The Porter stemming algorithm (or ‘Porter stemmer’) is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.
Input:-connected/connecting/connection/connections Output:-connect

The stop word removal algorithm:-To remove useless words from a search query string, in the Framework. These words, such as "the", "because" and "how", are considered stop words and can usually be safely ignored.

Input: I saw a cat and a horse.

Output: saw cat horse

6.4 Clustering & Its Processing

1. Sample results for General & specific Query

The system was assessed for a number of real-world queries; also analyzed the results obtained from our system with respect to certain characteristics of the input data. The queries are mainly categorized in four types such as

1 Ambiguous Query

2 General Query

3 Compound Query

4 People Name

The system was tested for all these queries & the result obtained is satisfactory. The figure 10 shows the clusters obtained for query “mouse” whereas the figure 11 shows the relevant pages under the cluster “Computer Mouse”.

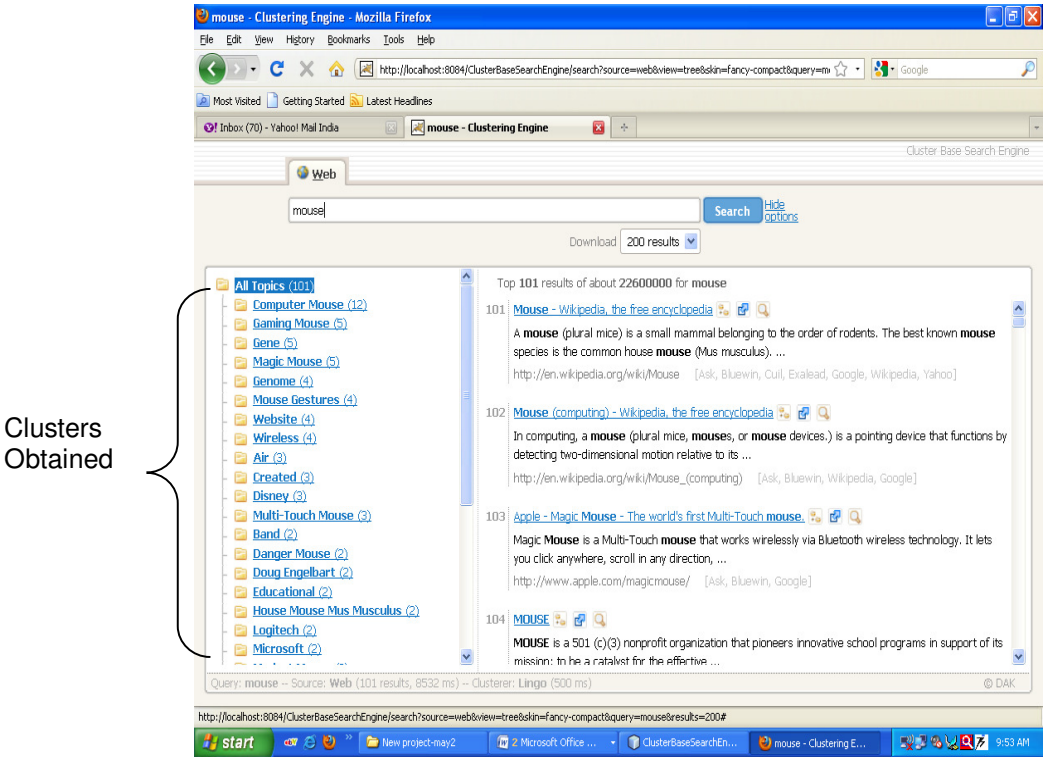

FIGURE 10: Clusters for a ambiguous query “mouse” & k=200 results

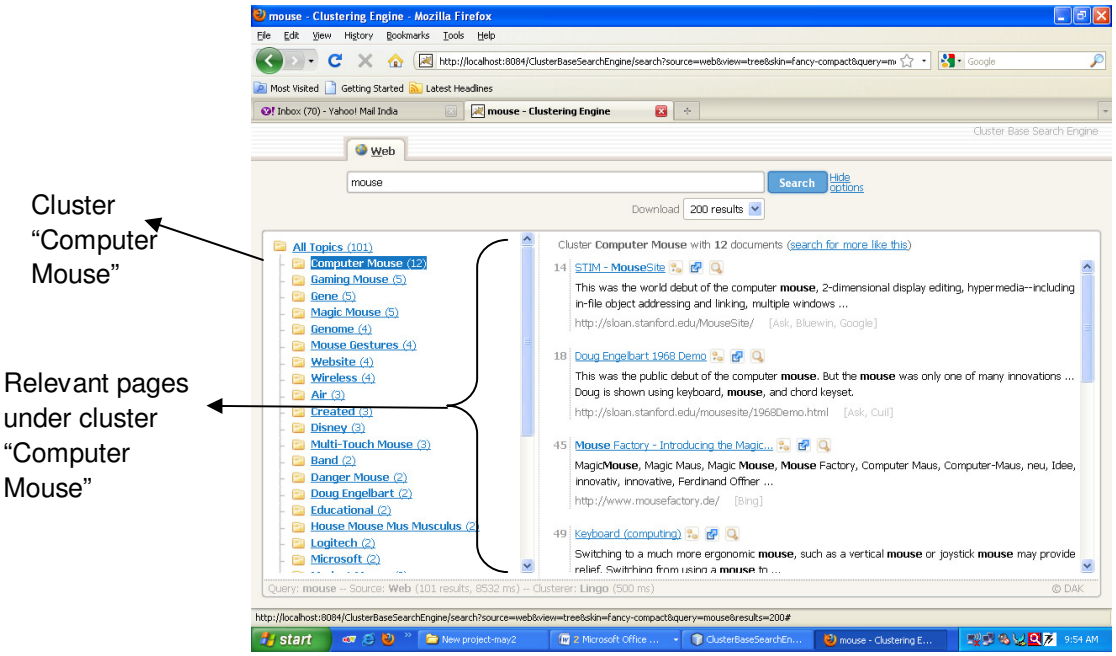

FIGURE 11: Relevant pages under the cluster "Computer Mouse"

From each type of query, some sample queries are taken for the testing of system. The Table 3 shows the name of query under each type, its clusters obtained & the processing time required. In the table the clusters are shown for top 200 results for each query. From the list of clusters obtained, only first 15 clusters are shown in the Table 3.

Query Type:-Ambiguous			
Sr. No	Name of Query	Clusters	Time
1	Mouse	Computer Mouse, Mickey Mouse, Website, Cells from Mouse, Gaming Mouse, Common Mouse, House Mouse, Technology, Gene, Graphics, Series, Apple, Magic Mouse, Support, Windows	844 ms
2	Tiger	Panthera tigris , Year of Tiger, Animal, News, Tiger Woods, Tiger Pictures, Website, Map, Tiger reserve, Art, Mac OS X, Tiger Airways, Wikipedia, Big Cats, China	906 ms
3	Java	Java Developers, Java Language, Java Platform, Download, Sun Microsystem,Java Tutorials, Java Applets, Java Virtual Machine, Java News, Java Resoursec,Community, Java Implementations, Java Runtime Environment, Java EE, Open Source, Learn Java,Class,Coffee	735 ms
4	Apple	Apple Computer, Mac OS X, Apple Products, Music, Developer, New Yark, iPod, Macintosh, Valley, AAPL Stock, Experience, Hardware & Software , Programs, Series, Tree	735 ms
5	Sahara	Western Sahara, Sahara Desert, Sahara Services, Map, Sahara Group, Design, Film Directed,India,Photos,Vegas Hotels, Website, Matthew McConaughey, Sahara Occidental, South Africa	719 ms

6	Apache	Apache Project, Open Source, Apache Software Foundation, Web Server, Apache Related, Apache Tribe, Acute Physiology & Chronic Health Evaluation, Apache License, Download, United States, Apache II Score, New Maxico, Reviews, Apache Maven, Attack Helicopter	891 ms
7	Jobs	Job Opportunities, Positions, Services, Market, Career Advice, Jobsuche, Research ,Switzerland Jobs, Books, Employee, Graduates, Post Your Resume, Career Opportunities, Employment Opportunities, People	735 ms
8	Tree	Tree of Life, Trees Used, Tree Structure, Family Tree, Forest, History, Tree of Life Web Project, Abstract, Data Structure, Green, Service, Art, Defining, Living, Maximum Likelihood, Music	875 ms
9	Ups	Company,Tracking,Shipping,UPS Freight, Uninterruptible Power Supply, Protect, APC, Press, Uninterruptible Power Systems, United Parcel Service, Air Freight, Browse UPS Portfolio UPS Solution Finder Solutions, Fields,Ireland,Package	265 ms
10	Jaguar	New Cars, Jaguar Panthera Onca, Jaguar XF, Jaguar XJ, Performance, Engine, Reviews, Google Buchsuche-Ergebnisseite, Land, Buy, Classic, Jacksonville Jaguars, Luxury, Big Cats, Jaguar E-Type	657 ms
11	Saturn	Planet Saturn, Rings Saturn, Saturn's Moons, Jupiter & Saturn, Solar System, Saturn Cars, News, View Saturn, Google Buchsuche-Ergebnisseite, Saturn Wikipedia, Saturn Corporation, Downloads, Fact, ION, Media Market	844 ms
12	Jordan	Michael Jordan, Air Jordan, Jordan Website, Photos, Jordan News, Jordan Born, December, MySpace, Hashemite Kingdom of Jordan, Woman, Middle East, New York, Research, Review, Wikipedia	640 ms
13	Quotes	Quotes & Quotations, Famous Quotes, Quotes Collection, Market Quotes, Quotes Saying, Love Quotes, Inspirational Quotes, Quotes of the Day, Trades & Quotes, Funny Quotes, Auto Insurance, Quote Topics, Compare Auto, Humor, Research, Life Insurance	672 ms
14	Matrix	Systems, Matrices, Method, Matrix Games, Analysis, Wikipedia, Extracellular Matrix, Human, Matrix Group, Revolutions, Applications, Download, International, Laurence Fishburne, Material, Power, Select	656 ms
15	Light	Light Source, Design, Seeing the light contact, Electric, Electromagnetic Radiation,News,Organic,Search,Website,Blue,Energy,Music,Point,Video	828 ms
Query Type:-General Terms			
16	Yellow pages	Search Yellow pages, Local Business Directory, White Pages, Local Business Listings, Phone numbers, Maps & Directions, Telephone Directory,Companies,Classifieds,International,People,Sourse,City Guides, Complete, Global	806 ms
17	Maps	World Maps, Google Maps, Driving Directions, Download Maps, Digital Maps, Image Map, Bing Maps, Historic, Yahoo, Genetic Linkage Maps,	859

		Germany Map, International, Live, Models, National Geographic, Code	ms
18	Health	Health news, Health medical, National Health, Health Fitness, Health & Medical News, Public Health, Centers, Diet, Children's Health, Diseases & Conditions, Health Articles, Health System, Health Information Resources, Health Magazine, Health Promotion	687 ms
19	Flower	Flowering plants, Send Flowers, Delivery Send Flowers, Flower Garden, Shop, Flower Show, International, Pictures, Wedding, Directed, Flower Power, Mothers Day, Protein, Research, Stem	718 ms
20	Music	MUSIC Videos, New MUSIC, MUSIC Song, World MUSIC, Internet, Country MUSIC, MUSIC Lyrics, Listen to MUSIC, MUSIC Games, Search, Performance, Electronic MUSIC, Album, Internet, Radio, MUSIC Festival	719 ms
21	Chat	Chat Rooms, Live Chat, Chat with people, Chat with friends, Community, Web Chat, Chat System, Games, Help, Blogs, News, Voice Chat, Chatrooms, Login, Windows	688 ms
22	Games	Puzzle Games, Flash Games, Download Games, Play Arcade, Fun Games, Racing Games, Alzheimer's Disease,Internet,Summer Olypics, Kids, Learning, Funny Games, News, Models, Publisher	672 ms
23	Radio	Radio Broadcasting, Music Radio, Internet Radio, News, Digital Radio, Talk Radio, Radio Stations Live, Streaming Radio, Public Radio, Local, Switzerland, BBC Radio, Radio Shows, Analysis, Programming	828 ms
24	Jokes	Blonde Jokes, Collection, Cartoons, Category, English, Games, Funny Pictures, Laugh, People, Clean Funny, Sex, German, Lots, Computer, Human	672 ms
25	Graphic design	Art design, Web design, Website design, Advertising, Graphic design Illustrator, Graphic design Portfolio, Graphic design Studio, Graphic design Schools, Typography, World of Graphic design, Graphic design Blog, Graphic design Program, Marketing, Process, Freelance Designers, Graphic design Agency	703 ms
26	Resume	RESUME Services, Cover letter, Writing Samples, Templates, RESUME Help, Sample Cover, Tips, Search, Graphic Design, Jobs Post, Portfolio, RESUME Builder, Website, Examples & Templates, Publishing	687 ms
27	Travel	Information & Travel, Cheap Hotels, World Travel, Travel Holiday, Travel Tourism, Travel Services, Destination Guide ,Car rentals, Time Travel, Airline Tickets, Switzerland Travel, Tours, Travel Blog, Travel Directory, Travel Discounts	671 ms
28	Time zones	Countries & Time Zones, Standard Time, Daylight Saving Time ,Time Map, World Clock, Time Difference, Current Local Time, UTC TIME, Time Zones are observed, United states, Date Time, Eastern, Page Time Zones, Time Zones	703 ms
29	World war 2	History of World, Second World, Timeline for World,WW2,World War 2,Germany in World, German World, Army, Day by Day, World War 2 Weapons, Europe after World, Video, Books, World War II Memorial, United States, Adolf Hitler	672 ms

30	Ford	Ford Motor Company, Ford Dealer, Ford Parts, Henry Ford, New Cars, Performance, Ford Fiesta, Reviews, Engine, Going, United States, Breast Cancer, Europe, Explorer, Limited	766 ms
Query Type:-Compound Query			
31	Travel to indonesia	Bali Indonesia, Indonesia Hotels, Indonesia Tours, Indonesia Islands, Travel Information & Travel Guide, Holiday in Indonesia, Indonesia, Jakarta Travel, Visit Indonesia, Travel Tips, Southeast Asia, Travel Warnings, City, Culture, History	797 ms
32	Photos of pets	Animal Photos, Pictures of Dogs, Pictures of Cats, Pet Photo Gallery, Pet News, Pets & People ,Featuring, Puppies, Digital, Kids, Life, Pets Like, Send, Contests, Health	657 ms
33	how to prepare for gate-11	Exam, GATE 2010,Time To Prepare, method, September 11,Attacks,Blog,City,July,Subjects,John,Present,Sample,Steps,Study	891 ms
34	the enemy of the state	Enemy of the state movie, Enemy of the state 1998,Review of Enemy of the state, Trial & Execution of Saddam Hussein, Video, Enemy of the state film, Gene Hackman , Enemy of the People, Lupe Fiasco, Jerry Bruckheimer, Directed by Tony Scott, Office, Public Enemy, Wolverine, Love story Mix tape, Declared, Enemy of the State Award, Making, Crimson Tide, Death	703 ms
35	topics related to thrust areas in computer engineering	Related Research, Computer Science & Engineering,University,Related Course, Education, Electrical & Computer Engineering Department, Mechanical Engineering, Civil Engineering, Management, New Thrust, Systems Engineering, School, Response	704 ms
36	To be or not to be	Google Buchsuche-Ergebnisseite, Games, Tobe T, Face book, Name Tobe, News, Reviews ,Seiten , Service , People, Play Games, Question, Search, Tobe, Hooper, Years	781 ms
37	New York times	City New York, New York times Company, New York Times Newspaper, Breaking News, New York Times Magazine, United States, New York Times Copyright, Readers, Privacy Policy, Report, New York Review of Books, Travel, Car, Daily Newspaper, Media	703 ms
38	making best use of existing resources	Practices, Available Resources, Decision Making, Programs, Software, Sources, Existing Work, Making more efficient use of Existing Resources,Difference,Health,Natural,Training,Land,Report,Schools	812 ms
39	Define skill & confidence	Confidence Definition,Learning,Skill Level,Self,Improve,Communication Skill, Skill Set, Social Skill, Apply Media, Leadership Skill,Game,Test,Coaching,Life,Measure,Facilitator	834 ms

40	giving wings to your future	Giving wings to Dreams, Future Plans, Life,Tip,Children's Future, Future Generation,Help,Save Your Search Strategies for Future Use, Wings to EAA,Spread,Idea,Program,Design,Detroit Red Wings,Mean,Thought	687 ms
Query Type:-People Name			
41	Pratibha Patil	Pratibha Patil News ,Woman Behind Woman, New Delhi, Woman, Governor of Rajasthan, Pratibha Patil Photos, Patil to visit, National, Pratibha Patil Pictures,rashtrapati Bhavan,Video,Female President,Manmohan Singh,Abdual Kalam,Presidental Candidate	969 ms
42	g a Patil	G.A., Pratibha Patil, Department, Georgia GA, Internal Medicine, Journal, Indian National Congress, R.T., S.L., A.Y., Nagana Gowda, Address & Phone Number, Education, PaR, Book	875 ms
43	d a kale	Kale Rang da Yaar,D.Kale,MP3 Download, Kale Rang de Paranda,ND,Dupatta Tera, Brassica Oleracea,Rk,Center,India,S.D.Kale,search Engine,Shazia Manzoor, Authors, High	781 ms
44	Kiran Bedi	India's First Woman, Indian Police Officer, Highest Ranking, Kiran Bedi Film, Kiran Bedi 2009,Woman,Indian Police Service IPS, New Delhi,www.kiranbedi.com-first & Highest Ranking Indian Woman,Foundation,Celebrated,Madam Sir, Reforms, Tihar Prison, Book	734 ms
45	Ujjwal Nikam	Mumbai Terror Attacks, Special Public Prosecutor in the 26/11,Terror Case, Videos, Terrorist, Accused, Latest News, Death Penalty Business, Media, Sabauddin's Acquittal, Search, Conduct Open Trial in 26/11 Case, Conviction, Witnesses	860 ms
46	Dr c d kane	Danity Kane, New CD, Albums,B C D E F G H,M.Kane,D.C,Movie,MySpace,Professor,RedPyramid,Feelgood,M.D,Medicine,Picture, Studio	719 ms
47	Andrew McCallum	Univerity, Author Andrew, People, Profile for Andrew McCallum, Conditional Random Fields, Text Classification,Kamal Nigam, University of Massachusetts Amherst, Labeled Data, Aron Culotta, LinkedIn, Music, Active, Graduate Student, MALLET	750 ms
48	William Stalling	Computer Organization & Architecture, Data Communication, EBook EBook, Network Security, Operating Systems, William Stallings Cryptography, Author William Stalling, Download Torrents,6th Edition,Bill, Rapidshare Search, Operating Systems Internals & Design Principles, Review, Network Security Essentials, Principles & Practice	766 ms
49	a a bachchan	Aishwarya Rai, Amitabh Bachchan's Movie, Bachchan's Videos,Yeh Kahan Aa Gaye Hum, Film Starring, Bachchan Family,Bollywood Star, Starring Abhishek Bachchan,Blog,Lata Mangeshkar,Role,Latest News, YouTube, Bachchan a Day, Film Directed,Harivansh Rai Bachchan	734 ms

50	Kelly Flanagan	J.Flanagan, Palm Desert Real Estate, Brigham Young University, Blog, VIDEO, Kelly Flanagan's Page, Computer Science Department, Engineering, Jobs, Member, Authors, Ian bannen, John Flanagan, Trace Collection, Elizabeth	750 ms

TABLE 3: First 15 Clusters for various queries

6. CONCLUSION & FUTURE WORK

Most of the traditional entity resolution techniques are feature-based similarity (FBS) methods, as they employ similarity functions that compare values of entity features (or attributes) in order to determine if two object descriptions co-refer. The estimation is that context-based and relationship-based algorithms always get better results than FBS because of entity-relationship graph generation, wherein the nodes represent the entities in dataset D and the edges represent the relationships among the entities. For any two entity representations, the co-reference decisions will be made not only based on the entity features, or information derived from the context, but also based on the inter entity relationships, that exist among the two representations. Because of this process, there will be improvement in the performance. The concentration will be made on study of the efficiency of the approach. Also there is probability that it will improve the result of clustering because disambiguation algorithm makes analysis of inter object connection. SVM classifier is used to determine whether the system is stable or not depending on the workload that is assigned for each & every machine treated as client machines also inform about the instability of system when excess load is there. During processing whole data is updated in database stored on server machine. Hence SVM done the work of load balancing over the network & provide a stable network where all requests get distributed equally. Due to this the efficiency of the system will be increased significantly.

7. REFERENCES

[1].D.V.Kalashnikov, S.Mehrotra, R.N.Turen and Z.Chen, "Web People Search via Connection Analysis" IEEE Transactions on Knowledge and data engg.Vol 20, No11, Novr 2008.

[2]. D.V. Kalashnikov, S. Mehrotra, Z. Chen, R. Nuray-Turan, and N.Ashish, "Disambiguation Algorithm for People Search on the Web," Proc. IEEE Int'l Conf. Data Eng. (ICDE '07), Apr. 2007.

[3]. M. F. Porter. *An algorithm for suffix stripping*. Program Vol. 14, no. 3, pp 130-137.

[4]. S. l. Osinski, J. Stefanowski, and D. Weiss. "Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition".

[5]. Z. Dong. "Towards Web Information Clustering". PhD thesis, Southeast University, Nanjing, China, 2002.

[6] S. l. Osinski. "An Algorithm for Clustering of Web Search Results". Master's thesis, Pozna´n University of Technology, Poland, 2003.

[7] G. Salton. "Automatic Text Processing — The Transformation, Analysis,and Retrieval of Information by Computer". Addison–Wesley, 1989.

[8] O. E. Zamir. "Clustering Web Documents: A Phrase-Based Method for Grouping SearchEngine Results". Doctoral Dissertation, University of Washington, 1999.

[9] J. Stefanowski and D. Weiss. "Web search results clustering in Polish- Advances in Soft Computing, Intelligent Information Processing and Web Mining", Proceedings of the International IIS: IIPWM´03 Conference,Zakopane, Poland, vol. 579 (XIV), 2003, pp. 209-22.