Intelligent Clustering Engine Solution for Desktop Usability

Prajakta Pawar
Pillai Institute of Information Technology, Engineering, Media Studies and Research

Sushopti Gawade
Pillai Institute of Information Technology, Engineering, Media Studies and Research

Sharvari Govilkar
Pillai Institute of Information Technology, Engineering, Media Studies and Research

ABSTRACT
ICE stands for Intelligent Clustering Engine. Term weighted similarity measure algorithm, phrase matching algorithm and Document Index Graph Based Clustering algorithm are the algorithms of the project 'Intelligent Clustering Engine Solution for Desktop Usability". The Engine is based on SAGH methodology. SAGH stands for Genetic Analytical System for Grouping Hypertext. The stages in SAGH are pre-processing, matrix formation, vectors, clustering, visualization, and output. Visualization enhancement is one of the Usability characteristic that has been enhanced within the proposed system Intelligent Clustering Engine. The paper focuses upon implementing the document clustering application over desktop to improvise usability.

Keywords
Intelligent Clustering Engine, Clustering, Usability, Carrot2, Document Clustering

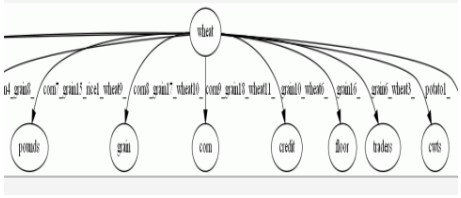
1. INTRODUCTION
In the recent years, computer and internet technology has been widely popularised and is now available to a wide variety of users. This has transformed traditional printed material into digital material and caused the search for information to move from libraries and catalogues to internet search engines using network links. By storing digital information in internet servers and sharing it by way of internet search engines, human knowledge is now undergoing a new revolution. Acquiring knowledge is no longer limited by geography, as a search engine can be shared and used by anyone, anywhere, anytime, using any internet browsing software. Intelligent Clustering Engine will be handling data content mining, data usage mining, document clustering and information retrieval process, which together form the base of system. The carrot2 system is the existing system taken from existing clustering search engines, and forms the base system for the Intelligent Clustering Engine, over desktop which is the proposed system. The information in the internet does not have the rigorous organization. There is a requirement of categorisation of search results. The global acceptance of the internet as a useful repository of knowledge makes us ignore the imperfect ordering of information in the vast and diversified structure of internet. Little time and effort to search for specific information represent very valuable aspect. There are some tools which strive for simplicity and agility of information When more and more queries are asked to engine, it takes more time for the engine to answer to the query. It results in more result overhead, with lesser throughput. Commercial clustering engine Grokker took a strategy whereby the search results clustering engine was delivered to the user as a downloadable desktop**.** This is the start of Desktop Clustering where both search results acquisition and clustering would use Client's machine resources, which are, network bandwidth and processor power. In this way, scalability issues and the resulting problems could be avoided.

2. LITERATURE REVIEW
Carlantonio et al have given the information about the existing web clustering engines. Then they have proposed the system called as intelligent clustering engine that removes the drawback of existing clustering engine and takes us through the entirely new approach of document clustering [1]. Meiyappan et al have constructed the base for the system (ICE) through the discovered SRCluster Engine [2]. Weiss et al have explained the Carrot2 clustering framework [3]. They have guided us by providing details of Carrot2 clustering engine. It aids by providing the Carrot2 clustering engine as a base for intelligent clustering engine [4]. Carpineto et al have explained the approach to web clustering engines and have given information regarding the survey of the clustering engines [5]. Li et al have given information about information architectures. They have also explained the approach in Carrot2 framework. They have discussed the process used in Carrot2 framework. Li has given the details of how the final output is obtained in the Carrot2 engine [9]. Sathya et al have given information about Clustering based on relevancy of documents and linking in retrieval system. When the web pages are clustered, a boost up factor is given to a web page based on the relevancy of content from title and summary [7]. Momin et al have given information about document clustering, which is an important tool for many Information Retrieval (IR) tasks. Document clustering techniques mostly rely on single term analysis of document data set. To achieve more accurate document clustering, more informative feature such as phrases are extracted [8]. Zhao Li et al have given information about measuring the similarity of short text snippets plays an important role in information retrieval and natural language processing [9]. Aggarwaal et al explain that clustering is a widely studied data mining problem in the text domains [10]. Wang et al have explained that Document clustering generates clusters from the whole document collection automatically and is used in many fields, including data mining and information retrieval. In the traditional vector space model, the unique words occurring in the document set are used as the features [12]. Huang et al have given information about clustering as a useful technique that organizes a large quantity of unordered text documents into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanisms [11]. Michael et al conclude that hierarchical clustering is often visualized as the better quality clustering approach. Website1 related to "http://www.aduna-software.com/technologies/autofocus/overview.view", gives information about Aduna based Autofocus Engine [14]. It is

an Autofocus Engine. Website2 "http://search.Carrot2.org /stable/search" gives information about the approach in Carrot2 framework [15]. Carrot's main priorities are to minimize the effort required to extend or alter the framework and to provide a wide variety of ready-to-use components. Website3http://nlp.stanford.edu/IR-book/html/ htmledition/single-link-and-complete-link-clustering-1.html [16]

3. IMPROVISATION IDENTIFIED

3.1 Visualization Improvement



Snapshot1: Cluster Visualization

Many documents which were to be represented in a haphazard form are now shown in a structured format. From this structure format, the user can locate the clusters and the documents in a relatively easy manner. See in Snapshot1. Output frame is created and the real visualization is implemented over the two dimensional view which is snapshot2. In this diagram the cluster names (links present in the output frame interface in the intelligent clustering engine) and the document data is displayed below it.

3.2 Documents organization

The documents are organized into clusters which are shown in the output above (Snapshot2). The documents are included in the clusters, which aids us to browse the documents with the help of SAGH technology [5].

3.3 Corpus Summarization

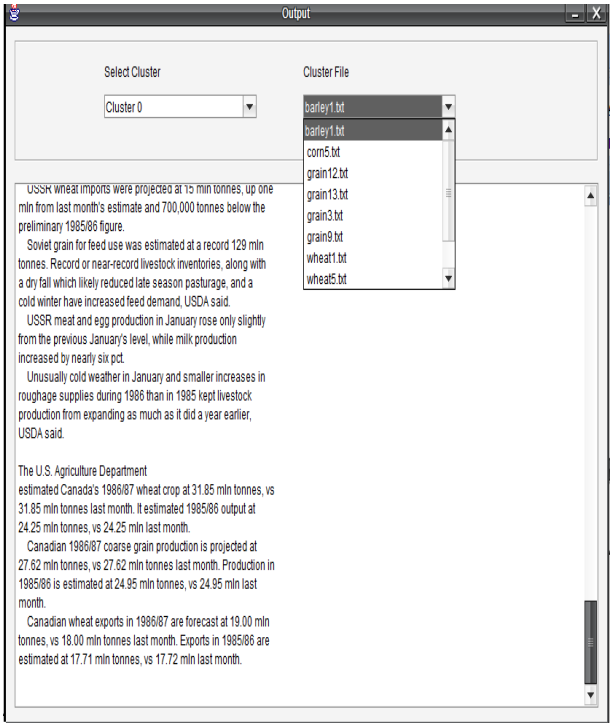
The corpus is the files Reuters dataset, the files in content and process generated at runtime, the text corpus generated and the features list. The input and the output generated in the form of the summary is the corpus summarization. Fast subtopic retrieval: If the documents that pertain to the same subtopic have been correctly placed within the same cluster and the user is able to choose the right path from the cluster label, such document can be accessed in logarithmic rather than linear time. Like 'wheat is the topic' and 'bushels' 'certs' are the subtopics. Like 'wheat is the topic' and 'bushels' 'certs' are the subtopics. The documents are placed in the same sub-topics and are accessed in the logarithmic and linear time. [10].

3.4 Document Classification

While clustering is inherently un-supervised learning method, it can be leveraged in order to improve the quality of the results in its supervised variant. Word-clusters can be used in order to improve the classification accuracy of supervised applications with the use of clustering techniques. [5] [8]

The list of stop words is defined and extracted from the data set of documents' data. Then we convert the filtered and extracted the document into the output files. See Figure1 above. In this figure it has been explained that the output of the system are ordered document vectors, Indexed document matrix, visualization and clusters.

The vector analysis says that the vectors are in linear form. We need to get the information about 'wheat'. We give the query in the engine. The output is features, document names, document clusters, cluster visualization, time accuracy graph. Then, we try to search a link between, the documents included in the cluster. Inside the output frame, we can find the link between the subtopic and the document.



Snapshot2: Output frame of Intelligent Clustering Engine

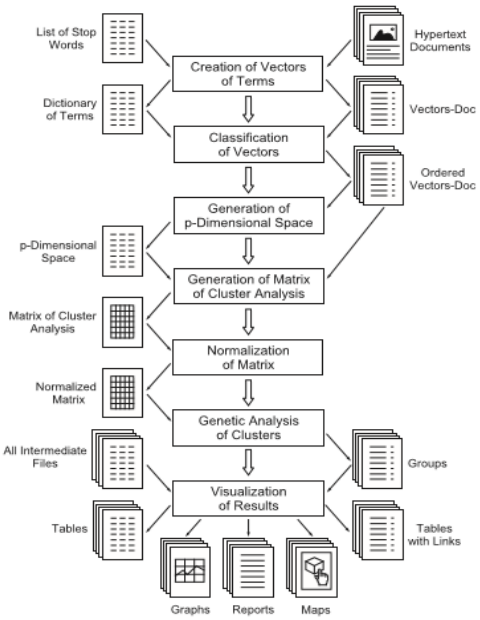
The two dimensional document matrix aids to identify the location of the term inside the text documents. The term document matrix defines the term frequency of individual term. The user reaches the right document in the finite amount of time, according to the time accuracy graph (example 32,500- 35,000 milliseconds for pre-processing in query 'wheat'). Intermediate files, groups or categories, time accuracy graph, reports, maps, are the output files [1].

4. METHODOLOGY

The methodology is defined by the programming logic used to generate results. Methodology consists of 3 phases input, processing and output. Input is the term query, processing is the Intelligent Clustering Engine, output is the clusters. Methodology used is SAGH methodology [1].

The process and the content folders store intermediate files at the time of pre-processing. The data is extracted and put into categories called clusters.

Each cluster consists of multiple documents. Text corpus summarization is implemented when user clicks upon the document name, example 'barley1.txt'.



Snapshot3: SAGH Methodology

5. ALGORITHM

5.1 Term weighted Similarity Measure

Algorithm [9]

Steps

1. Let Rn(x) be the set of top n documents returned by a search engine when using x as the query.
2. For each document in di ∈ Rn(x), construct the term vector vi with TF×IDF and truncate each vector vi to include its m highest weighted terms.
3. Let C(x) be the center of the normalized vectors vi

$$C(x) = \frac{1}{n}\sum_{i=0}^{k} vi/||vi|| \qquad (1)$$

4. For each element Ki ε C(x), suppose i*t is the term corresponding to Ki. Construct the term vector

$$WC(x) = ki \times \frac{E(ti)}{\sqrt{D(ti)}} \qquad (2)$$

5. Let $QE(x)$ be the normalization of the center WC (x):

$$QE(x) = \frac{WC(x)}{||WC(x)||} \qquad (3)$$

So, the similarity of two short snippets can be obtained by calculating the inner-product $(x) \times QE(y)$. The term weighted similarity measure is calculated from equations (1), (2), (3).

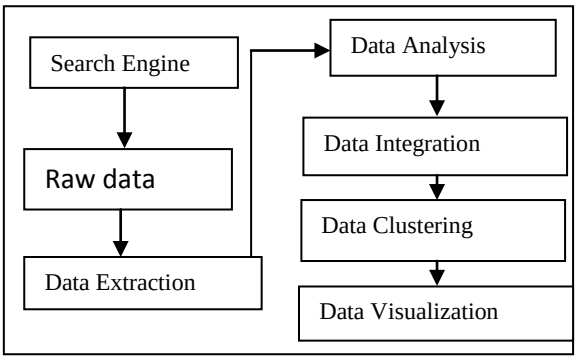


Figure: ICE – Block Diagram

5.2 Phrase Matching Algorithm [3]

Steps

1. di ← next document
2. M ← Empty List {list of matching phrase}
3. for each sentence sij in di do
4. v1 ← tij1 {first word in sij}
5. if (v1 is not in Gi-1) then
6. add v1 to Gi-1
7. end if
8. for each term tijk belong to sij, k = 2, …s, lij
9. vk← tijk, vk-1← tij(k-1), ek← (vk-1, vk)
10. if (vk is not in Gi-1) then
11. add vk to Gi-1
12. else
13. if ek is an edge in Gi-1
14. Retrieve the list of document table entries from
15. edge ek's table.
16. Extend previous matching phrases in M for
17. phrases that continue along edge ek.
18. Add new matching phrases to M.
19. else
20. add edge ek to Gi-1
21. end if
22. end if
23. update sentence path in vk-1 and vk
24. end for
25. end for
26. Gi← Gi-1
27. Output matching phrase list M

5.3 DIGBC Algorithm [8]

Steps

1. di ← new document
2. find out common phrases between di all existing clusters

3. compute similarity of document to all existing clusters.
4. for j = 0 to nclusters // nclusters ← number of existing // clusters (initially 0)
5. if (sim[di, cj] > threshold) then
6. add di to cj
7. modify document tables accordingly
8. end if dd
9. end for
10. if di is not assigned to any existing cluster
11. add di to a new cluster
12. increment nclusters
13. modify document tables accordingly.
14. end if

6. RESULTS

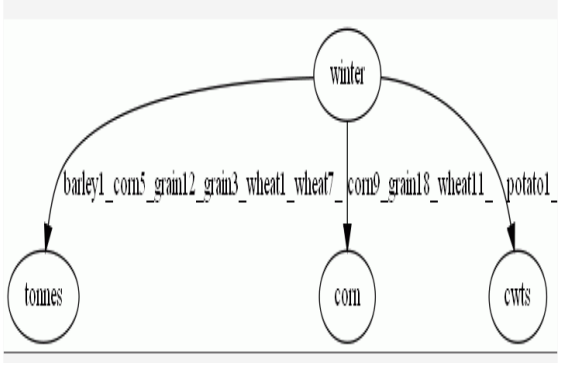
Result Analysis is performed with Carrot2 as Existing System and ICE as the Proposed System. In this process of analyzing result, a query is put like "wheat" is taken.

Table: Comparison between ICE and Carrot2. ICE gives less and accurate clusters whereas Carrot2 gives many and inaccurate queries

Clustering Engine	Clusters	Documents
ICE	10/38	30/218
Carrot 2	28/38	92/218

Reuters Dataset

It is the dataset consisting of text files. They are present in 45 categories. Data is queried from this dataset.



Snapshot4: Cluster for query 'winter' in ICE

7. CONCLUSION

Intelligent Clustering engine enhances the clustering methodology, in terms of information retrieval. Searching alone does not satisfy the user requirement, in terms of information availability, information visualization. There are specifically two kinds of output obtained after implementing the project, which are document categories and summarised text. This summarised text is the collection of all the text data related to the query term 'wheat'. The overall approach is to improvise the user satisfaction in terms of information usage. Every required information about "wheat" as query gives all existing information about quantity of wheat, quality of wheat, rate of wheat, yearly productivity of wheat and related products all required supportive information that exists over desktop; that is done on one place. This has been achieved by clustering enhancement through SAGH system over desktop. Visualization improvement is done. The query can be used for rest of the categories as wheat. Clustered Grouping the data in a systematic manner such that user retrieves the usable information within minimum effort, is been implemented.

8. FUTURE SCOPE

Work in the future would be extending the system for the clustering of files of type image. It can be extended to internet as 'Application as a Service'. Also, it would be allowed clustering to any type of indexed document in this application.

9. 9. REFERENCES

[1] Intelligent Clustering Engine: A clustering gadget for Google Desktop, Authors- Carlantonio, Osiek, Xerox, Costa, Journal- ESWA, 2012

[2] SRCluster Web Clustering Engine based on Wikipedia, Authors-Meiyappan, Iyengar and Kannan, Journal-International Journal of Advanced Science and Technology, Vol. 39, February, 2012

[3] A Survey of Web Clustering Engines, Authors-Carpineto, Stanislaw, Romano, Weiss, Journal- ACM Computing Surveys, Vol. 41, No. 3, Article 17, July 2009

[4] Carrot2 Clustering Framework, Authors- Weiss, Poland and Osi´nski, Journal-EKMA, July 2005

[5] A Document Clustering Approach for Search Engines, Authors- Liang, Huei Ho, Yang and Chiang, Journal-IEEE International Conference on Systems, Man, and Cybernetics October 2006

[6] An Efficient Algorithm for Clustering Search Engine Results, Authors-Zhang, Pang, Xie, and Wu, IEEE 2006

[7] Link Based K-Means Clustering Algorithm for Information Retrieval, Authors- Sathya, Jayanthi, Basker, TN, India, Journal - IEEE, 2011

[8] Web Document Clustering Using Document Index Graph, Authors- Momin, Kulkarni, Chaudhari, Journal-IEEE, 2011

[9] An improved measuring similarity for short text snippets and its application in Clustering Search Engine, Authors-Li, Peng, Peng, Jia, Journal- Seventh International Conference on Machine Learning and Cybernetics, Kunming, July 2008

[10] A Survey of text clustering algorithms , Authors-Aggarwal, Zhai, Watson Research Center, University of Illinois, Journal- IEEE, 1984- 2012

[11] Similarity Measures for Text Document Clustering, Author-Anna Huang, Department of Computer Science, University of Waikato, 1998- 2007

[12] Document Clustering with Semantic Analysis, Authors-Wang, Julia Hodges, Department of Computer Science &

Engineering, Journal - 39th Hawaii International Conference on System Sciences, 2006

[13] A Comparison of Document Clustering Techniques, Authors- Steinbach, Karypis, Kumar, Department of Computer Science and Engineering, University of Minnesota

[14] Website1: Aduna–AutoFocus, Link- http://www.aduna-software.com/technologies/autofocus/overview.view,Visited in July, 2009, Carrot2 Search (2011). Website2: Carrot2 clustering engine, link- http://search.Carrot2 .org/stable/search, visited September 2011

[15] Website3: http://nlp.stanford.edu/IR-book/html/htmledition/single-link-and-complete-link-clustering-1.html Single Linkage clustering, Complete Linkage Clustering