ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

WEB MINING: DOCUMENT FILTERING IN E-COMMERCE USING CLUSTERING

Ashok K. Panda[1], Dhiren K. Sahu[2], S.N.Dehuri[3], M.R.Patra[4]

Associate Professor, Dept. of Computer Sc., MITS Engg College, Rayagada, Odisha, India[1]

PG Student of IT, Dept. of IT & Communication, Utkal University, Bhubaneswar, Odisha, India[2]

Associate Professor, Dept. of Systems Engg. Ajou University, Suwon, South Korea[3]

Reader, Dept. of Computer Science, Berhampur University, Berhampur, Odisha, India[4]

Abstract: Document filtering is probably the most challenging task in the Web. Giving a prominent search result by filtering the document is a measure issue. Semantic similarity and large document clustering is the most difficult task as the web data has a lot of redundancy like outliers, missing values etc, data prepossessing is very much necessary. Search results produced by social search engine (web search) give more visibility to the content created. This paper focuses on semantic similarity measure, the F-measure for large document clustering. Document filtering is a task to retrieve documents relevant to user's profile. Generally, filtering systems calculate the similarity between the profile and each incoming document for retrieving documents with similarity having higher threshold value. With the increased use of the Internet and the World Wide Web, E-commerce transaction is growing rapidly. Therefore, finding useful patterns and rules of users' behaviours has become the critical issue for E-commerce and is used to tailor e-commerce services to meet the customers' need successfully. In this paper, we highlight the ArteCM clustering algorithm and implemented it which provides better results for document filtering for retrieving most relevant documents in E-commerce transaction.

Keywords: document filtering, e-commerce, clustering, world wide web, ArteCM clustering algorithm.

I. INTRODUCTION

The birth of internet is really a gift to the mankind. In the recent years the growth and popularity of the internet has increased to such an extent that every person knows about it and uses it for various purposes. Some people use internet to know new things, while others use it as a means of entertainment. The use of internet is not only limited to the entertainment but it can also be used to conduct research related to work or study, get latest news etc. Now a day's people uses internet for E-commerce. Popularity of E-commerce is so high that it's very difficult to manage web stuff. With each passing movement millions of web pages are added to this internet.

The implementation of search engines on the internet made the process of searching some of the topics very easy. Querying the search engine for any particular topic would retrieve the results from the internet and those results are then presented to the users. But since there are many pages on the internet the results obtained by the search engines are also vast. It becomes really difficult for the user to get the particular page from the search engine. If it happened that the Page Rank of the particular page is high then it can be found on the first page of the search engine results, else it can be found at the end of the results. This results in the loss of time for the users as they had to spend the time looking for the particular required page. To overcome the drawbacks of the search technique, it is necessary that the search results are clustered.

Clustering will help to group the similar pages together and the dissimilar pages are not grouped Presenting this grouped results to the user will help the users to get all the related pages to their query and also will reduce the time spent by them in searching the related page. Presently there are various recommendations and techniques to cluster the web pages. This paper proposes one of the clustering systems which clusters the web pages by taking in the user query.

E-commerce : Electronic commerce [18], commonly known as E-commerce, is the buying and selling of product or service over electronic systems such as the Internet and other computer networks. Electronic commerce draws on such technologies as electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web at least at one point in the transaction's life-cycle, although it may encompass a wider range of technologies such as e-mail, mobile devices and telephones as well.

Web mining : Web mining is a very hot research topic which combines two of the activated research areas: Data Mining and World Wide Web. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Although there exists quite some confusion about the Web mining, the most recognized approach is to categorize Web mining into three areas: Web content mining, Web structure mining, and Web usage mining. Web content mining focuses on the discovery/retrieval of the useful information from the Web contents/data/documents, while the Web structure mining emphasizes to the discovery of how to model the underlying link structures of the Web. The distinction between these two categories isn't a very clear sometimes. Web usage mining is relative independent, but not isolated, which mainly describes the techniques that discover the user's usage pattern and try to predict the user's behaviours. There are three phases in web mining [17] as given below:

1. Infrastructure
   - Crawling The Web
   - Web Search and Information Retrieval
2. Learning
   a. Similarity and Clustering
   b. Supervised Learning
   c. Semi Supervised Learning
3. Application

Document Filtering : Document filtering is a task which monitors a flow of incoming documents, and selects those which the systems regard as relevant to the user's interest. Many document filtering systems use a similarity-based method to retrieve documents. The user's interest is expressed within the system as a profile. The similarity between the profile and each incoming document is calculated, and documents with similarities higher than a preset threshold are retrieved. Retrieved documents are sent to the user, who returns a relevance feedback to the system. This feedback information is used to update the profile for the upcoming flow of new documents.

Clustering : The process of forming the group of similar items is known as clustering. The process of clustering can be used in various fields such data clustering, document clustering, web clustering, etc. Given a certain data points, consider some of the data point as the centroid and calculate the distances of other points with respect to the chosen centroid. Putting the certain threshold on to the maximum distance, the data points which are within the threshold will gel with the respective centroids and the clusters are formed. The total number of clusters formed, depends upon the initial number of centroids selected for clustering. There are various types of clustering algorithms with some advantages and disadvantages. The main types of clustering algorithms are Partitional, Hierarchical and Density-based clustering algorithms.

Web Page : A web page [20] is a web document or other web resource that is suitable for the World Wide Web and can be accessed through a web browser and displayed on a monitor or mobile device. This information is usually in HTML or XHTML format, and may provide navigation to other web pages via hypertext links. Web pages frequently subsume other resources such as style sheets, scripts and images into their final presentation. Web pages may be retrieved from a local computer or from a remote web server. The web server may restrict access only to a private network, e.g. a corporate intranet, or it may publish pages on the World Wide Web. Web pages are requested and served from web servers using Hypertext Transfer Protocol (HTTP). Web pages may consist of files of static text and other web content stored within the web server's file system (static web pages), or may be constructed by server-side software when they are requested (dynamic web pages). Client-side scripting can make web pages more responsive to user input once on the client browser.

Web Content : Web content [21] is the textual, visual or aural content that is encountered as part of the user experience on websites. It may include, among other things: text, images, sounds, videos and animations. In Information Architecture for the World Wide Web, Lou Rosenfeld and Peter Morville write, "We define content broadly as 'the stuff in your Web site.' This may include documents, data, applications, e-services, images, audio and video files, personal Web pages, archived e-mail messages, and more. And we include future stuff as well as present stuff.

HTML head **:** In a web page there are two section , head and body. Body section contains the content that displays in the web page but head section doesn't have any role for the content. But there some information like title and Meta information that can be useful for the content is concern.

HTML Script Tag **:**HTML script tags are using for many client side interaction, form validation, animation and to give rich functionality to a web page

HTML Style Tag **:**HTML style tag using to add eye catching style, colour and positioning to the web page through internal as well as external style sheet

HTML No Script **:**Some of the old browser doesn't support script like java script or if JavaScript has been disabled in a web page the script tag will not work and scripts will display in the web page to avoid this no script tag are there to use

*HTML Comment***:** Comment always adds additional information about a particular subject. Like wise in html also there are comments to describe the code functionality.

Stop Words : While calculating the frequency of the terms appearing the document, care is so taken that, the prepositions, conjunctions, adverbs, verbs are avoided. These are the terms form stop words. It is most likely that considering the stop words in the process of clustering will definitely lead us to wrong results. The reason to discard these stop words is, because of the frequency of these stop words in a document is very high. If these are not discarded, they will play role in calculating the inverse document frequency which will directly affect the cosine similarity index thereby effecting the clustering results. Stop words that are present in frequency in the documents are as exemplified below:

"able,about,above,according,accordingly,across,actually,after,afterwards,again,against,ain't,all,allow,allows,almost,alone,along,already,also,although,always,am,among,amongst,an,and,another,any,anybody,anyhow,anyone,anything,anyway,anyways,anywhere,apart,appear,appreciate,appropriate,are,aren't,around,as,aside,ask,asking,associated,at,available,away,awfully,be,became,because,become,becomes,becoming,been,before,beforehand,behind,being,believe,below,beside,besides,best,better,between,beyond,both,brief,but,by,c'mon,c's,came,can,can't,cannot,cant,cause,causes,certain,certainly,changes,clearly,co,com,come,comes,concerning,consequently,consider,considering,contain,containing,contains,corresponding,could,couldn't,course,currently,definitely,described,despite,did,didn't,different,do,does,doesn't,doing,don't,done,down,downwards,during,each,edu,eg,eight,either,else,elsewhere,enough,entirely,especially,et,etc,even,ever,every,everybody,everyone,everything,everywhere,ex,exactly,example,except,far,few,fifth,first,five,followed,following,follows,for,former,formerly,forth,four,from,further, etc.

II. RELATED WORKS

An In Carullo et al paper [1] the ArteCM Algorithm describes the similarity measures on short documents which defines the speed & time of retrieving relevant short documents. Alexander Budanitsky and Graeme Hirst [2] focus on similarity or semantic distance in WordNet which were compared by examining their performance in a real-word spelling correction system. It determines the degree of semantic similarity, relatedness between two lexically expressed concepts. The paper by - Keiichiro Hoashi [3] proposes the use of a non-relevant information profile in order to retrieve more relevant documents without excessive retrieval of non-relevant documents. The object of this profile is to reject non-relevant documents which are similar to documents mistakenly retrieved in the past flow of documents. The paper [4] by- Chi Lang Ngo, [4] in their paper describing clustering based on rough sets , proposes a Tolerance Rough Set Clustering algorithm for web search results and implementation of the proposed solution within an open-source framework. The paper by Elizabeth D. Foused[5] shows implementation and testing of the SFC order as means for semantically representing the content of texts for the purpose of delimiting document set with a high likelihood of containing all those relevant to an individual query proving the results as promising. Nicola cancedda in his paper [6] describes the algorithm implemented by KerMIT consortium for its participation in the TREC 2001 filtering track consortium using a liner SVM with an innovative threshold section mechanism for the adaptive task using both a second order perceptron with uneven margin. Courtney Corley [7] presents a knowledge-based method for measuring the semantic similarity of texts. And introduced a method that combines word to-word similarity metrics into a text-to text metric showing the method outperforming the traditional text similarity metrics based on lexical matching. Eric Gaussier [8] proposed an online algorithm to learn category specific thresholds in a multi-class environment where a document can belong to more than one class.

In his paper Richard M. Paper [9] on “Advanced Decision Systems Division” is conducting a program of research to investigate machine learning techniques that can automatically construct probabilistic structures from a training set of documents with respect to a single target filtering concept, or a set of related concepts. Abbattista F in the paper [10] presents a personalization component that uses supervised machine learning to induce a classifier able to discriminate between interesting and uninteresting items making use of textual annotations usually describing the products in E-commerce. In the paper of B. Piwowarski [11] ,the author proposes an approach to build a subspace representation for documents is a first step towards the development of a quantum-based model for Information Retrieval(IR) validating to apply into the adaptive document filtering task. Inderjit Dhillon [12] suggests techniques for feature or term selection along with a number of clustering strategies , significantly reducing the dimension of the vector space model. In his paper Oren Zamir [13] introduce a novel clustering methods that intersect the documents in a cluster to determine the set of words (or phrases) shared by all the documents in the cluster more faster than other standard techniques. Byoung-Tak Zhang [14] presented a method that acquires re-inforcement signals automatically by estimating user’s implicit feedback from direct observations of browsing behaviours. The proposed learning method showed superior performance in information quality and adaptation speed to user preferences in online filtering. In the paper David A. Hull [15] systematically compare combination strategies in the context of document filtering using queries from the tipster reference corpus. Jamine Challan in the paper [16] describes a new statistical document filtering system called in -Route, for filtering effectiveness and efficiency that arise with such a system, and showed experiments with various solutions.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

III. THE PROBLEM STUDY

This section discusses the heuristics approach to fast, online document clustering based on domain specific similarity measures and describes the ArteCM algorithm [5]. Let D be the domain of documents d and D a given document collection, we define a normalized document similarity measure S:

$$S : \mathbb{D} \times \mathbb{D} \rightarrow [0;1] \qquad (1)$$

a normalized similarity measure S between a set of documents and a single document:

$$\bar{S} : 2^{\mathbb{D}} \times \mathbb{D} \rightarrow [0;1] \qquad (2)$$

$$\bar{S}(\bar{D}, \hat{d}) = \frac{\sum_{i=1}^{|\bar{D}|} S(d_i, \hat{d})}{|\bar{D}|} \qquad (3)$$

The ArteCM Clustering Algorithm : Requirement : Choose threshold parameter ε, Choose threshold parameter η, be $\hat{C}$ a growing set of elements C_i from 2^D.

```
1  for all dj € D do
2      m = argmaxi S(Ci, dj)
3      if S(Cm,dj) ≥ ε then
4          if S(Cm,dj) ≤ η then
5              Cm = Cm U { dj}
6          endif
7      else
8          C^new = {dj}
9          Ĉ = Ĉ U C^new
10             endif
11         Endfor
```

A. A threshold parameter ε € [0; 1] that defines the minimum similarity $S(C_i,d_j)$ document d_j must have in order to be assigned to cluster C_i.

B. A threshold parameter η € [ε;1] that defines the maximum similarity $\hat{S}(C_i,d_j)$ a document d_j must have to contribute to the definition of cluster C_i.

C. The two parameters play a fundamental role in the cluster growing process; the ε parameter directly controls the granularity of the document collection partitioning; while η parameter controls the number of elements considered in similarity computations, having a strong impact on overall speed. A standard similarity measure S_D - the Dice coefficient [10] with binary term weights, appropriate for our context and defined as:

$$S_D(d_i, d_j) = \frac{2C}{A + B} \qquad (4)$$

Where C is the number of common terms between di and dj , A and B are the number of terms of di and dj , respectively. A novel similarity measure S_T aimed to better fit the nature of the short documents domain where a "weighted" similarity measure can be easily applied in which common terms contribute with different weights in function of their typology (numbers, words, special chars, : : :).

$$S_T(d_i, d_j) = \sum_{r=1}^{|R|} \alpha_r \cdot \frac{2C_r}{A_r + B_r} \qquad (5)$$

Such that $\sum_{r=1}^{|R|} \alpha = 1$ and where $F=\{f_1,\ldots\ldots,f_{|F|}\}$ is the set of term types and C_r is the number of common terms of type f_r between d_i and d_j, A_r and B_r are the number of terms of type f_r in d_i and d_j respectively.

Discussions :

Evaluation Metrics : The evaluation phase takes into account cluster quality and speed, since we want to investigate fast clustering algorithms that can be applied on the fly on a collection of documents. In the Information Retrieval and Document Analysis field a widely accepted evaluation metric is the FMeasure (F1), as an harmonic mean between Precision and Recall [3] indexes. Given a collection of documents $D = \{d_1;\ldots; d_N\}$ and a list of labels $L = \{l_1;\ldots; l_M\}$ where M <= N we define the truth cluster set $C = \{C_1; : : : ;C_M\}$ where $Ci = \{fd_j$: the label of document d_j is $l_i\}$. If a single cluster C_i and an approximation of it $\hat{} C_j$ are considered, F_c is the F1 computed considering C_i as the set of relevant documents and $\hat{} C_j$ as the set of retrieved documents.

$$F1(\hat{C}, C) = \frac{\sum_{j=1}^{|C|} |Cj| \cdot F1_c(\hat{C}_{\text{argmax}_i (F1(\hat{C}i, Cj))}, C_j)}{\sum_j |C_j|} \qquad (6)$$

Being $\hat{} C$ a cluster set computed by an algorithm and following [4] the F1 within two cluster sets can be computed in terms of F1c: for each truth cluster the one with higher F1 is selected and then the weighted mean of F1 within all the cluster set is computed. The K-Means iterative algorithm is able to provide quite good results, even though the need to

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

know something about the number of needed clusters can be a limit in the web domain. Computing time though linear in the document collection size, can increase unexpectedly.

IV. IMPLEMENTATION .

Web Document Filtering using Clustering :

Here the ArteCM algorithm has been implemented for web document clustering. The main objective is to get the whole content of the web site and then pre-process the content and clean it. A web page contains various unnecessary information like head, script, style, no script, comment including stop words. To get actually and useful content we need to clean the data. Get the title of the website, Meta keywords and Meta description of the web page , then we calculate the word count then assign the site to a token id , then comparing the site with another site and find the common words between them. Then calculate the dice coefficient similarity measure. Then this process will continue for all the document di that belongs to the document collection , D. The whole steps of operation as detailed below.
The implementation process is carried out in the steps as follows :

a). Begins with fetching the content form the web site by unique web URL.
b). Then we pre-process and clean the content to get accurate result.
c). Then finding the word count of different portion of the content as unique word and whole word.
d). In the next step we compare the two sites finding their common word count and similarity measure.
e). Finally putting the dice-coefficient similarity measure in ArteCM algorithm we implement argmax operation it to create different clusters.

The whole process starting form fetching to similarity measure has been implemented using most popular and very fast growing server side programming language PHP and MySQL database.

Site Meta Data Table

The Table1 (Meta Table) contains the short name of the all type of word count do to lack of space we have to apply this technique to adjust the title of web pages 8 different type of word count. This data has been used bellow .

TABLE I
SITE META DATA

HEADING	SHORT NAME
All total word count	ATWC
All total unique word count	ATUWC
Filtered word count	FWC
Filtered Unique word count	FUWC
Tagged word count	TGWC
Title word count	TWC
Meta keywords count	MKC
Meta description count	MDC

There are more then 100+ E-commerce site taken in consideration in the experiment. It is not possible to show all the 100+ site information. The word collection and the words of only 8 website have been displayed here for understanding. This table contains the word count of the different part of the website in 8 different base.

TABLE 2
SITE INFORMATION

	SITE URL	ATWC	ATUWC	FWC	FUWC	TGWC	TWC	MKC	MDC
1	http://www.themobilestore.in	367	235	317	203	17	25	0	0
2	http://www.mobile9.com	108	94	69	62	6	9	0	24
3	http://www.virginmobile.in	180	141	111	95	13	15	0	47
4	http://www.bestylish.com	45	39	26	25	2	13	0	43
5	http://www.yuvastyle.com	515	320	415	280	82	19	0	22
6	http://www.metroshoes.net	141	106	108	86	17	13	0	25
7	http://www.yebhi.com	158	102	82	63	15	16	76	23
8	http://www.fashionara.com	244	186	175	146	19	12	0	52

Here in the Fig.2 contains 8 unique web page word count on 8 basis as described below :

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

All total word count (ATWC): All total word of count of a site is the total words before filtering the stop word but after cleaning the html head, script, style, no script and comment and html tag information from the fetched content.
All total unique word count (ATUWC) : These are the total unique word from the all total word because there may exist some duplicate word in the total word collection. After removing the duplicate word from all the word we'll get the all total unique word count (ATUWC)
Filtered Word Count (FWC): There might be some stop word like a, an, the some, etc... in the word collection so after filtering the stop words from the total word collection we are getting the filtered word count.
Filtered Unique word count : **A**fter getting the filtered word there might be some duplicated by removing the duplicate words we can get the filtered unique words count.
Tagged Word Count : Tagged word is really very interesting collection of words. Those words having high frequencies in the total word that'll be included in the tagged word count. These are collected from the filtered word.
Title Word Count : Every web page has <title></title> tag. This tag contains the title information of the web page. The total word count of the title is the title word count.
Meta Keywords Count : In the Head section of web page there is Meta keywords information that helps for identify the web page while searching in search engine. Total Meta keywords are the Meta keyword count. *Meta Description count :* Meta Description is the short description of the web page that present in side the head section of the web page. The total word present in the description of the Meta is the Meta description count.

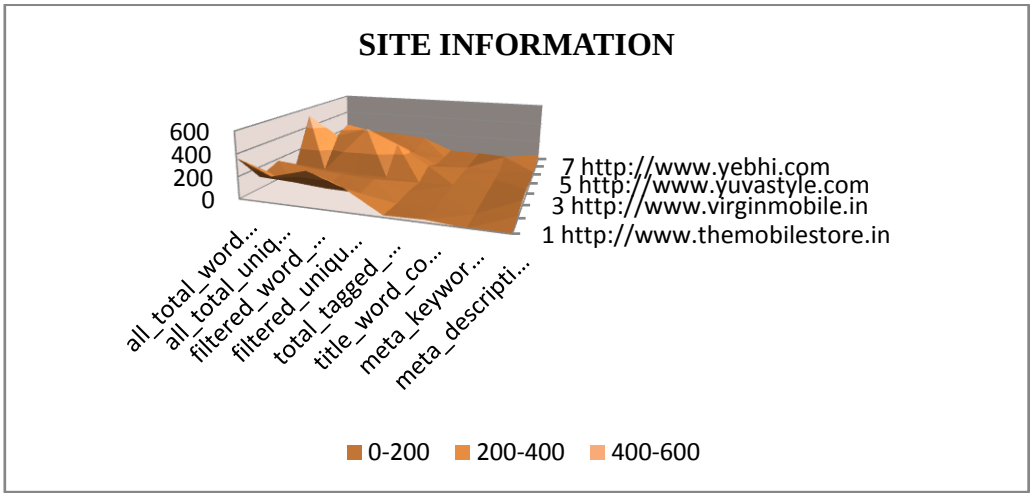

Fig. 2.Site Information Chart

Site word count inofrmation :
In the experiment we have taken more then 100 E-commerce website but it is not possible to give all the web site word count and similarity measure information. Here we present 8 unique URL and their word count on the basis of different criteria as given bellow.

Similarity Measure Meta Data Table

TABLE 3
SIMILARITY MEASURE META DATA

HEADING	SHORT NAME
Common All Total Word Count	CATWC
Common All Total Word Similarity Measure	CATWSM
Common All Total Unique Word Count	CATUWC
Common All Total Unique Word Similarity Measure	CATUWSM
Common Filtered Word Count	CFWC
Common Filtered Word Similarity Measure	CFWSM
Common Filtered Unique Word Count	CFUWC
Common Filtered Unique Word Similarity Measure	CFUWSM
Common Tagged Word Count	CTGWC
Common Tagged Word Similarity Measure	CTGWSM
Common Title Word Count	CTWC
Common Title Word Similarity Measure	CTWSM
Common Meta Keywords Count	CMKC

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

Common Meta Keywords Similarity Measure	CMKSM
Common Meta Description Count	CMDC
Common Meta Description Similarity Measure	CMDSM
Average Similarity Measure	ASM

In the Table.3 it contains the Meta data for the finding the common word count and similarity measure between sites. As the labels are very large, a short labelled Meta data table has been created. After getting the site information the site is assigned by a token id to identify uniquely. Then by comparing two site we'll get some common word. After getting the common word we can find the Dice Coefficient Similarity Measure by using the formula, equation-(4) , in S_D , where C is the number of common terms between di and dj , A and B are the number of terms of di and dj , respectively. Here we are finding the similarity measure of CATWSM, CATUWSM, CFWSM, CFUWSM, CTGWSM, CTWSM, CMKSM, CMDSM, then get the Average Similarity Measure(ASM) .

Site Comparison Common Word Table

TABLE 4
SITE COMPARISION COMMON WORD

SL NO	SITE 1	SITE 2	CATWC	CATUWC	CFWC	CFUWC	CTGWC	CTWC	CMKC	CMDC
1	2	1	17	14	7	7	0	1	0	0
2	3	1	24	13	6	6	0	7	0	0
3	3	2	29	19	6	5	1	1	0	10
4	4	1	15	10	1	1	0	5	0	0
5	4	2	7	5	1	1	0	1	0	2
6	4	3	9	6	1	1	0	1	0	7
7	5	1	62	15	4	3	0	4	0	0
8	5	2	34	17	4	4	0	1	0	1
9	5	3	49	22	9	7	0	2	0	3
10	5	4	48	15	5	5	0	8	0	7
11	6	1	30	16	17	11	0	6	0	0
12	6	2	9	8	5	5	0	1	0	1
13	6	3	5	4	2	2	0	2	0	3
14	6	4	16	8	5	5	0	11	0	8
15	6	5	37	19	21	12	2	8	0	8
16	7	1	45	9	7	2	0	6	0	0
17	7	2	14	12	3	3	0	1	0	1
18	7	3	25	19	8	6	0	1	0	6
19	7	4	36	6	1	1	0	9	0	6
20	7	5	52	20	8	8	1	7	0	8
21	7	6	43	8	10	5	1	5	0	5
22	8	1	40	20	12	8	0	3	0	0
23	8	2	21	11	1	1	0	0	0	4
24	8	3	30	16	1	1	0	1	0	5
25	8	4	15	6	1	1	0	6	0	13
26	8	5	52	26	13	10	4	5	0	10
27	8	6	38	21	25	15	3	5	0	8
28	8	7	31	13	2	2	1	6	0	3

Fig.3 and Fig.4 describes the Table.4 and Table.5 information respectively. In Table.4 it shows the common word count of all the 8 sites. There 28 records present each record have 8 attribute. Fig.3 showing the graph representation of

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

the Table.4 which clearly understandable. Similarly Table.5 contains the similarity measure graph contains the similarity measure by comparing two sites.

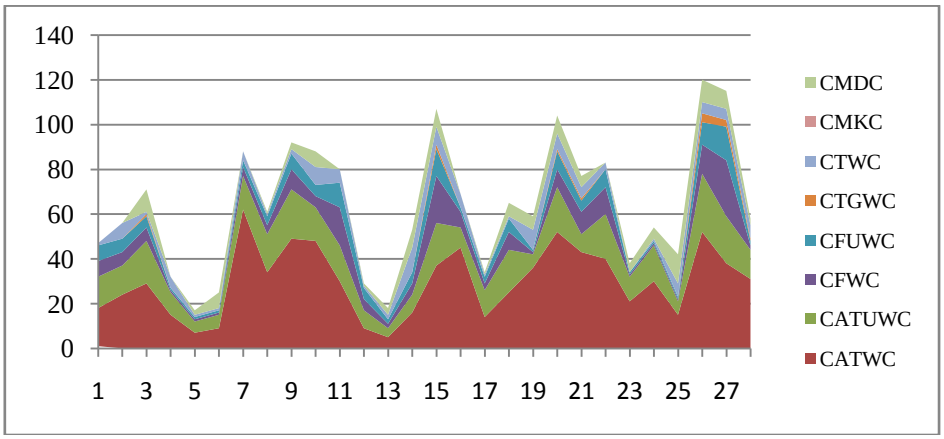

Fig. 3 . Site Comparison Common Word

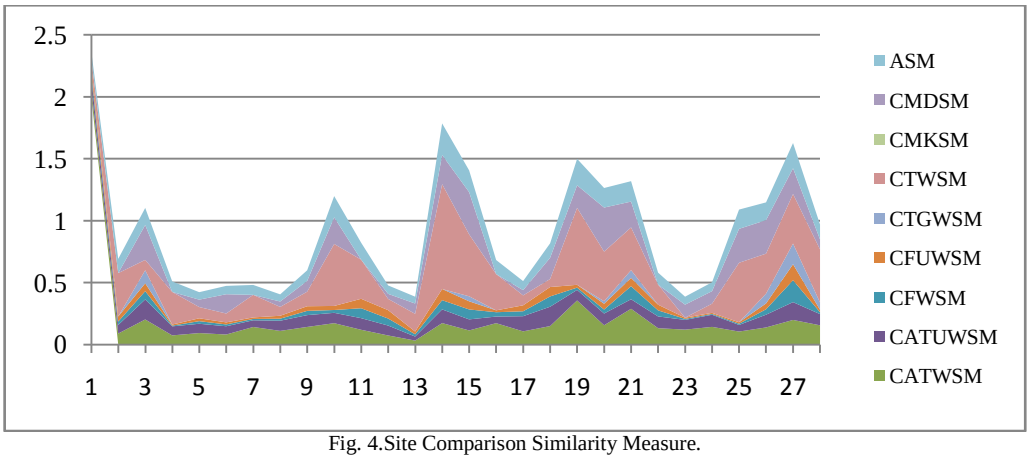

Fig. 4.Site Comparison Similarity Measure.

Fig.4 show graph for Table.5 contains the graph for similarity measure of all the comparisons between all the 8 site..Finally it calculates the average similarity form the all 8 base in Table.5. in Table 5 the similarity measure has been calculated by taking the common word count from the Table.4.

Site Comparison Similarity Measure :

TABLE 5
SITE COMPARISION SIMILARITY MEASURE

	SITE 1	SITE 2	CATWSM	CATUWSM	CFWSM	CFUWSM	CTGWSM	CTWSM	CMKSM	CMDSM	ASM
1	2	1	0.0715789	0.0851064	0.0362694	0.0528302	0	0.0588235	0	0	0.0609217
2	3	1	0.0877514	0.0691489	0.0280374	0.0402685	0	0.35	0	0	0.115041
3	3	2	0.201389	0.161702	0.0666667	0.0636943	0.105263	0.0833333	0	0.28169	0.137677
4	4	1	0.0728155	0.0729927	0.0058309	0.00877193	0	0.263158	0	0	0.0847138
5	4	2	0.0915033	0.075188	0.0210526	0.0229885	0	0.0909091	0	0.0597015	0.0602238
6	4	3	0.08	0.0666667	0.0145985	0.0166667	0	0.0714286	0	0.155556	0.067486
7	5	1	0.14059	0.0540541	0.010929	0.0124224	0	0.181818	0	0	0.0799626
8	5	2	0.109149	0.0821256	0.0165289	0.0233918	0	0.0714286	0	0.0434783	0.0576837
9	5	3	0.141007	0.0954447	0.0342205	0.0373333	0	0.117647	0	0.0869565	0.0854349
10	5	4	0.171429	0.0835655	0.0226757	0.0327869	0	0.5	0	0.215385	0.170974
11	6	1	0.11811	0.0938416	0.08	0.0761246	0	0.315789	0	0	0.136773
12	6	2	0.0722892	0.08	0.0564972	0.0675676	0	0.0909091	0	0.0408163	0.0680132

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

13	6	3	0.0311526	0.0323887	0.0182648	0.0220994	0	0.142857	0	0.0833333	0.055016
14	6	4	0.172043	0.110345	0.0746269	0.0900901	0	0.846154	0	0.235294	0.254759
15	6	5	0.112805	0.0892019	0.0803059	0.0655738	0.040404	0.5	0	0.340426	0.175531
16	7	1	0.171429	0.0534125	0.0350877	0.0150376	0	0.292683	0	0	0.11353
17	7	2	0.105263	0.122449	0.0397351	0.048	0	0.08	0	0.0425532	0.0730001
18	7	3	0.147929	0.156379	0.0829016	0.0759494	0	0.0645161	0	0.171429	0.116517
19	7	4	0.35468	0.0851064	0.0185185	0.0227273	0	0.62069	0	0.181818	0.213923
20	7	5	0.154532	0.0947867	0.0321932	0.0466472	0.0206186	0.4	0	0.355556	0.157762
21	7	6	0.287625	0.0769231	0.105263	0.0671141	0.0625	0.344828	0	0.208333	0.164655
22	8	1	0.130933	0.0950119	0.0487805	0.0458453	0	0.162162	0	0	0.0965465
23	8	2	0.119318	0.0785714	0.00819672	0.00961538	0	0	0	0.105263	0.064193
24	8	3	0.141509	0.0978593	0.00699301	0.00829876	0	0.0740741	0	0.10101	0.0716241
25	8	4	0.103806	0.0533333	0.00995025	0.0116959	0	0.48	0	0.273684	0.155412
26	8	5	0.137022	0.102767	0.0440678	0.0469484	0.0792079	0.322581	0	0.27027	0.143266
27	8	6	0.197403	0.143836	0.176678	0.12931	0.166667	0.4	0	0.207792	0.203098
28	8	7	0.154229	0.0902778	0.0155642	0.0191388	0.0588235	0.428571	0	0.08	0.120944

Steps for Document Clustering Implementation

There are various steps as given bellow that helps to create the document cluster :

1. Data Acquisition

Fetch Whole Content of The Web Page
Fetch Title Of The Web Page
Fetch Meta Keywords Of The Web Page
Fetch Meta Description Of The Web Page

2. Data Cleaning

Remove The Java Script Information From Page Content
Remove The Style sheet Information From Page Content
Remove The No Script Information From Page Content
Remove The Head Information From Page Content
Strip Multi-Line Comments Including C data from Page Content
Remove The Html Tags From Page Content
Remove The Stop Words From Page Content

3. Data Processing

Get All Words Of The Web Page
Get All Unique Words Of The Web Page
Get Filtered Words Of The Web Page
Get Filtered Unique Words Of The Web Page
Get Tagged Words Of The Web Page
Get Title Words Of The Web Page
Get Meta Keywords Of The Web Page
Get Meta Description Words Of The Web Page

4. Find Word Count

Get All Words Count Of The Web Page
Get All Unique Words Count Of The Web Page
Get Filtered Words Count Of The Web Page
Get Unique Filtered Words Count Of The Web Page
Get Tagged Words Count Of The Web Page
Get Title Words Count Of The Web Page
Get Meta Keywords Words Count Of The Web Page
Get Meta Description Words Count Of The Web Page
Create A Token Assign The Web Page To A Token

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

5. Finding The Dice Coefficient Similarity Measure

Get All Words Similarity Measure Of The Web Page
Get All Unique Words Similarity Measure Of The Web Page
Get Filtered Words Similarity Measure Of The Web Page

Get Unique Filtered Words Similarity Measure of The Web Page
Get Title Words Similarity Measure of The Web Page
Get Meta Keywords Words Similarity Measure of The Web Page
Get Meta Description Words Similarity Measure of The Web Page
Get Average Similarity Measure of The Web Page

6. Finding F1 Measure

In statistics, the F1 score (also F-score or F-measure) is a measure [22] of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

Implementing ArteCM Algorithm:

In section 3 , The problem study of ArteCM algorithm is described and implementing on it , from the similarity measure information , we get the cluster of documents using the F1 measure score in order to get the proper document cluster. The Document Cluster that is found is presented in the Table. 6.

Through the Fig.5 we found the final document cluster. There are three clusters we got after the whole process. The 8 unique web page has been clustered into three cluster ladled as "online Mobile Store", "Online Fashion Store" and Online "Footwear Store". In the first cluster there are 3 pages and 2nd cluster contains 4 pages while in the 3rd cluster we got only 1 page.

TABLE 6
FINAL DOCUMENT CLUSTER FORMED

SL N	SITE URL	LABLE
1	http://www.themobilestore.in	Online Mobile Store
2	http://www.mobile9.com	Online Mobile Store
3	http://www.virginmobile.in	Online Mobile Store
4	http://www.bestylish.com	Online Fashion Store
5	http://www.yuvastyle.com	Online Fashion Store
6	http://www.metroshoes.net	Online Footwear tore
7	http://www.yebhi.com	Online Fashion Store
8	http://www.fashionara.com	Online Fashion Store

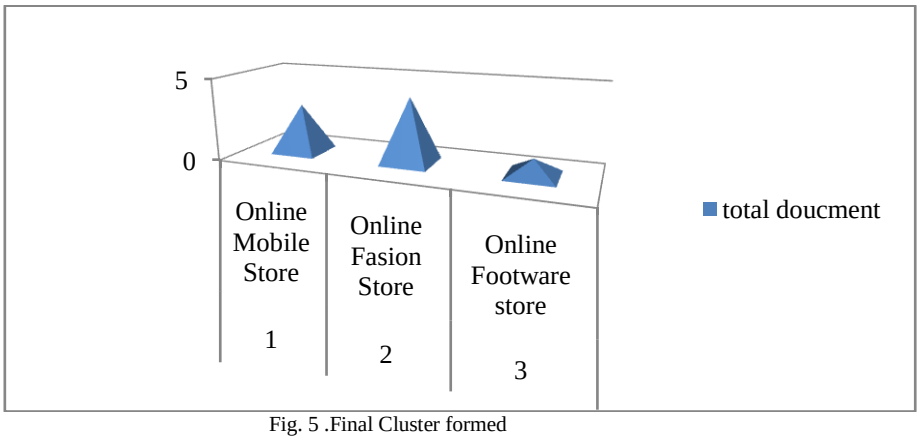

Fig. 5 .Final Cluster formed

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology
Vol. 2, Issue 6, June 2013

V. CONCLUSION AND FUTURE DIRECTION

In this paper we proved how the clustering method is very useful and effective. By adding some more levels in data cleaning we get more pure and the most relevant data, which ultimately helps to find better results. Taking title, Meta keyword and Meta description into account, it is more convenient in checking the similarity and relative document. Creating the tagged word is another advantage of this clustering method. Taking unique word is also a better measure. Web document is very complex and large. The clustering method hence create clusters from large and complex web documents as depicted in this paper. More than 100 E-commerce sites have been taken in implementing the algorithm for this experiment. It shows a good result with less time.

There is a number of future scope on this work. Researchers are encouraged to introduce some new similarity measure(s) for broader (large) document clustering. Computational applications typically require relatedness rather than just similarity. As the web document is very complex, it needs to be more clean and pure so that in future, addition of some new data cleaning techniques will provide more accurate and valued results

REFERENCES

[1] Moreno Carullo, Elisabetta Binaghi and Ignazio Gallo, Nicola Lamberti "Clustering of Short Commercial Documents for the Web", 19th International Conference IEEE Pattern Recognition, ICPR 2008.
[2] Alexander Budanitsky and Graeme Hirst "Semantic distance inWordNet: An experimental, application-oriented evaluation of five measures" , Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, June 2001.
[3] Keiichiro Hoashi Kazunori Matsumoto Naomi Inoue Kazuo Hashimoto "Document Filtering Method Using Non-Relevant Information Profile", SIGIR ' In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp.176-183, 2000.
[4] Chi Lang Ngo, Hung Son Nguyen, "A method of web search result clustering based on rough sets", WI '05 Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 673-679 , 2005.
[5] Elizabeth D. Liddy, Woojin Paik, " Semantic Information from Machine Readable Dictionary", Workshop On Very Large Corpora: Academic And Industrial Perspectives , 1993.
[6] Nicoloa Cancedda, Nicolces Branchi, Alex Conconi, Claudio Gentile, "Kernal Method for document filtering", Department of Commerce, National Institute of Standards and Technology, pp.19 – 22, 2002.
[7] Courtney Corley and Rada Mihalcea, "Measuring the Semantic Similarity of Texts", EMSEE '05 Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pp 13-18, 2005.
[8] Eric Gaussier, Ali Mustafa Qamar, and Vincent Bodinier, "Working Notes for the InFile Campaign : Online Document Filtering Using 1 Nearest Neighbor", CLEF 2008 Workshop, Aarhus, Denmark, pp.17-19, 2008.
[9] Richard M. Tong, Lee A. Appelbaum, "Machine learning techniques for document filtering", Association for Computational Linguistics Stroudsburg, PA, USA , HLT '93 Proceedings of the workshop on Human Language Technology, pp. 383-383, 1993.
[10] Abbattista F., Degemmis, Fanizzi N., Licchelli O., Lops P. Semeraro G., and Zambetta, F., "Learning User Profiles for Content-Based Filtering in e-Commerce", AIIA Conference, Siena, 2002.
[11] B. Piwowarski, I. Frommholz, Y. Moshfeghi, M. Lalmas, and .J. Van Rijsbergen, "Filtering Documents with Subspaces", Springer-Verlag Berlin, Heidelberg,ECIR'2010, Proceedings of the 32nd European conference on Advances in Information Retrieval, pp. 615-618, 2010.
[12] Inderjit Dhillon, Jacob Kogan, Charles Nicholas, "Feature Selection and Document Clustering", ACM New York, NY, USA SAC '11 Proceedings of the 2011, ACM Symposium on Applied Computing, pp 1143-1150, 2011.
[13] Oren Zamir and Oren Etzioni, Omid Madani, Richard M. Karp, "Fast and Intuitive Clustering of Web Documents", KDD, pp. 287-290 1997.
[14] Byoung-Tak Zhang and Young- Woo Seo, "Personalized Web-Document Filtering Using Reinforcement Learning", Applied Artificial Intelligence Volume:15, pp:665-685 2001.
[15] David A. Hull, Jan O .Pedersen, Hinrichs Shutze, "Method Combination for document filtering", ACM New York, NY, USA, SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 279-287 1996.
[16] Jamine Challan, "Document filtering with inference networks", ACM New York, NY, USA SIGIR '96 Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp 262-269 , 1996.
[17] Chakrabarti Soumen (2003), " Mining the Web Discovering Knowledge From Hypertext Data", Morgan Kaufmann Publication, 2005.
[18] http://en.wikipedia.org/wiki/Electronic_commerce
[19] http://en.wikipedia.org/wiki/Web_document
[20] http://en.wikipedia.org/wiki/Web_page
[21] http://en.wikipedia.org/wiki/Web_content
[22] http://en.wikipedia.org/wiki/F1_score