A Hierarchical Clustering Approach for the Creation of a Simple Semantic Web Application

Apeh, Ayo I.[1], Olatunde, Olabiyisi S.[2], Owolabi, Olumide[1]
1.Computer Centre, University of Abuja, Abuja, Nigeria
2.Department of Computer Science, Ladoke Akintola University of Technology, Ogbomosho, Nigeria

Abstract

The goal of the Semantic Web is to develop enabling standards and technologies designed to receive more exact results when searching for information, and to help machines understand more information on the Web so that they can support richer discovery, data integration and navigation. This can be achieved if there is a common vocabulary for a set of domains. Information is published using standard vocabulary. This study explores the processes of creating a taxonomy for a set of journal articles using hierarchical clustering algorithm. 100 journal articles that cut across different fields were downloaded from the internet. These served as sample data. These journal articles were serialized, stemmed and tokenized. Term frequency was calculated for each journal article. Some representative terms were selected from each journal article and similarity matrix was generated for the entire journal articles. Complete hierarchical clustering was used to create a cluster of the articles. JavaTree view program was used to view the dendrogram of the cluster. It was observed that the articles cluster around their subject, subject area, field of study, area of application, journal type, author, place of case study. This demonstrated that journal articles have properties on a taxonomy, could be created as a basis for a semantic web.

Keywords: Semantic web, clustering, taxonomy, similarity, document collection.

1. Introduction

The World Wide Web contains huge amounts of information created by many different organizations, communities and individuals for many different reasons. Users of the Web can easily access this information by specifying URL addresses, searching, and following links to find other related resources. The simplicity of usage is a key aspect that has made the Web so popular. However, the simplicity of the current web has a price. It is very easy, with all that is available, to get lost, or discover irrelevant or unrelated information For instance, if we search for something as simple as research papers written by a person named "Eric Miller" we will find all kinds of other information starting from Web diaries or phonebooks that mention "Eric" and/or "Miller" somewhere. Similar problems arise if we search for resources about "Marja", as "Marja" could equally well refer to a first name of a person, or to a berry in Finnish (Riitta and Miller., 2001).

The Semantic Web is all about developing enabling standards and technologies to help machines understand more information on the Web so that they can support richer discovery, data integration, navigation, and automation of tasks. With Semantic Web we not only receive more exact results when searching for information, but also know when we can integrate information from different sources, know what information to compare, and can provide all kinds of automated services in different domains from future home and digital libraries to electronic business and health services (Bernes-Lee, 2001).

Currently, the World Wide Web is based mainly on documents written in Hypertext Markup Language (HTML). With this setup, humans are capable of using the Web to carry out tasks such as reserving a library book, ordering for goods etc. However, a computer cannot accomplish the same tasks without human direction because web pages are designed to be read by people, not machines. Considering this limitation, the dream of the web has not been completely realized.

It is against this background and limitations that the idea of the semantic web was postulated by Berners-Lee when he expressed a vision of the future web as follows:

I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A ‘Semantic Web’, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The ‘intelligent agents’ people have touted for ages will finally materialize (Bernes-Lee,1999)

The semantic web is a vision of information that is understandable by computers, so that they can perform more of the tedious work involved in finding, sharing and combining information on the web. It is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the contents of the web. The data on the present web are not only heterogeneous but have different structures aside the fact that they are presented in HTML format, which is more of text format. This study, therefore, sought to develop taxonomy for some set of documents as a gateway to a semantic web application. Organizing the set of documents in a hierarchy according to how we judge them to be related should facilitate intelligent search, such that, once a document is located we can always link other documents most closely related to it. This will also form the basis for a web of documents, such that new documents will always be inserted in the most appropriate places alongside related documents.

2. Related Work.

At its core, the semantic web comprises a set of design principles collaborative working groups, and a variety of enabling technologies. Some elements of the semantic web are expressed as prospective future possibilities that are yet to be implemented or realized. Other elements of the semantic web are expressed in formal specifications (Herman, 2008). Some of these include Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, Notation 3 (N3), Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain. The semantic web has also been described as a component of Web 3.0 (Berners-Lee et al, 2006)

With HTML and a tool to render it (perhaps web browser software, perhaps another user agent), one can create and present a page that lists items for sale. The HTML of this catalog page can make simple, document-level assertions such as "this document's title is 'Widget Superstore'". But there is no capability within the HTML itself to assert unambiguously that, for example, item number X586172 is an Acme Gizmo with a retail price of N1000, or that it is a consumer product. Rather, HTML can only say that the span of text "X586172" is something that should be positioned near "Acme Gizmo" and " N1000", etc. There is no way to say "this is a catalog" or even to establish that "Acme Gizmo" is a kind of title or that "N1000" is a price. There is also no way to express that these pieces of information are bound together in describing a discrete item, distinct from other items perhaps listed on the page.

Semantic HTML refers to the traditional HTML practice of markup following intention, rather than specifying layout details directly. For example, the use of <em> denoting "emphasis" rather than <i>, which specifies italics. Layout details are left up to the browser, in combination with Cascading Style Sheets. But this practice falls short of specifying the semantics of objects such as items for sale or prices. Microformats represent unofficial attempts to extend HTML syntax to create machine-readable semantic markup about objects such as retail stores and items for sale (W3C, 2009)

The Semantic Web takes the solution further. It involves publishing in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML). HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts. Berners-Lee calls the resulting network of Linked Data the Giant Global Graph, in contrast to the HTML-based World Wide Web.

These technologies are combined in order to provide descriptions that supplement or replace the content of Web documents. Thus, content may manifest as descriptive data stored in Web-accessible databases, or as markup within documents (particularly, in Extensible HTML (XHTML) interspersed with XML, or, more often, purely in XML, with layout/rendering cues stored separately). The machine-readable descriptions enable content managers to add meaning to the content, i.e. to describe the structure of the knowledge we have about that content. In this way, a machine can process knowledge itself, instead of text, using processes similar to human deductive reasoning and inference, thereby obtaining more meaningful results and facilitating automated information gathering and research by computers.

3. Methodology

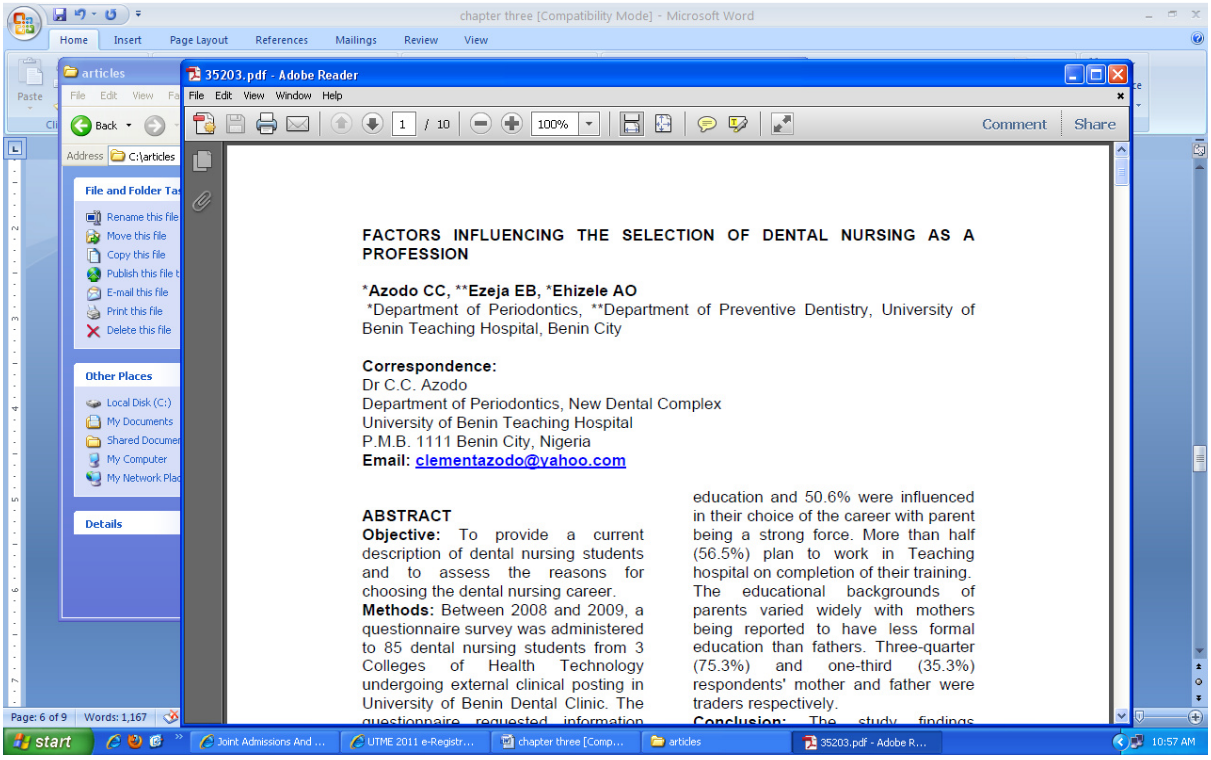
An attempt is made in this work to realize the goal of the semantic web by creating a structure that ties together research publications on the web in such a way that any collection of papers can be linked on the basis of similarity for the purpose of searching. If the papers are linked and categorized, then a search can always begin from the most promising point in the collection and will be sure to return the most relevant papers for any search criterion. New papers being added to the collection will also be done according to the existing categories, or lead to the creation of new categories, in order to continue to make for the most efficient search possible.

The methodology employed to realize the set goal is as follows:

Using a collection of 100 set of journal articles, the articles are serialized to remove document formats, leaving only the plain texts of the documents. The articles are then tokenized to get individual words or strings, which are then stemmed to the root words. Stemming enables us to get the same word for strings that are variants of the same word. For example, the words ‘stem’, ‘stems’, ‘stemming’ and similar words should all reduce to the string - ‘stem’. To identify descriptive terms for each document, term frequencies are computed for each article and representative terms are selected. These representative terms are then used to compute inter-document similarity values and a similarity matrix. Using hierarchical clustering algorithm, the articles are clustered to create taxonomy for the documents. Java tree view then is used to create a dendrogram of the clusters.

Table 1: A partial list of downloaded journal articles

Doc1	Contribution of Information Channels to Adoption of Aquaculture Technologies Among Fish Farmers in Anambra State and Implication For Training
Doc2	Profitability, Inputs Elasticities and Resource-Use Efficiency in Small Scale Cowpea Production in Niger State, Nigeria
Doc3	Performance of Mapping-Grade Gps Receivers in Southeastern Forest Conditions
Doc4	A Psychological Perspective On God-Belief As A Source of Well-Being and Meaning
Doc5	Factors Affecting Job Satisfaction of Rangers in Yankari Game Reserve, Bauchi, Nigeria
Doc6	Perception of Farmers About Profitability of Vegetable Gardening Enterprise in Ahiazu Mbaise Local Government Area of Imo State, Nigeria
Doc7	Economics of Small-Scale Palm Oil Processing in Ikwerre and Etche Local Government Areas of Rivers State, Nigeria.
Doc8	Effect of Water Harvesting Methods, Nitrogen-Phosphorus Fertilizer and Variety On Leaf Tissue N, and P, and Soil Moisture Content of Date Palm
Doc9	Effect of Water Harvesting Methods, Nitrogen and Phosphorus Fertilizer On Leaf Length of Different Date Palm (Phoenix D-) Varieties
Doc10	Assessment of The Frequency of Ict Tools Usage By Agricultural Extension Agents in Imo State, Nigeria
Doc11	Economic Analysis of The Effect of Organo-Mineral Fertilizer On Tomato Yield Component in Humid Forest Zone of Nigeria
Doc12	Antimicrobial Profile of Moringa Oleifera Lam. Extracts Against Some Food – Borne Microorganisms
Doc13	Evaluation of Different Morphotypes of Mango (Mangifera Indica L.) for Use As Rootstock in Seedlings Production
Doc14	Serodynamics of Treponema Pallidum in Serum of Pregnant Women in Benin City
Doc15	Effects of Soil Types and Enhanced Nutrient Levels On The Productivity of Earthworm (Eudrilius Eugeniae, Kinberg)
Doc16	Effects of Processing On The Mineral Content, Proximate Composition and Phytochemical Factors of The Seeds of Bauhiniamonandra (Kurz)



FACTORS INFLUENCING THE SELECTION OF DENTAL NURSING AS A PROFESSION

***Azodo CC, **Ezeja EB, *Ehizele AO**
*Department of Periodontics, **Department of Preventive Dentistry, University of Benin Teaching Hospital, Benin City

Correspondence:
Dr C.C. Azodo
Department of Periodontics, New Dental Complex
University of Benin Teaching Hospital
P.M.B. 1111 Benin City, Nigeria
Email: clementazodo@yahoo.com

ABSTRACT
Objective: To provide a current description of dental nursing students and to assess the reasons for choosing the dental nursing career.
Methods: Between 2008 and 2009, a questionnaire survey was administered to 85 dental nursing students from 3 Colleges of Health Technology undergoing external clinical posting in University of Benin Dental Clinic. The questionnaire requested information

education and 50.6% were influenced in their choice of the career with parent being a strong force. More than half (56.5%) plan to work in Teaching hospital on completion of their training.
The educational backgrounds of parents varied widely with mothers being reported to have less formal education than fathers. Three-quarter (75.3%) and one-third (35.3%) respondents' mother and father were traders respectively.
Conclusion: The study findings

Fig. 1: A pdf view of one of the articles

Data Serialization

The journal articles were published with different text and page formatting. To allow for homogeneity, the articles were converted to plain text where all text and page formatting were removed.

Tokenization

The text of the different journal articles were tokenized into strings called tokens, that is, the individual words constituting the text file were extracted. This was achieved with the use of a C++ program tagged Stack_Token. The program takes input file and produces substrings (tokens) and the frequency of occurrence of each word term) in the article. This was done for all the 100 journal articles. Stop words, such as 'is', 'a', 'to', and so on, in the text documents were removed leaving only significant terms which are relevant to the context of the document.

Stemming

In order to get unique words and avoid duplication of terms, the terms were stemmed to their root using the Porter stemming algorithm (Porteer, 2000). The Porter stemming algorithm reduces words based on their morphology. For examples, the following terms "computing", "computer", "compute", will be stemmed to "comput". The frequencies for all the words that reduce to the same root were also summed.

Selection of terms

Representative terms were selected for each document using the frequencies of occurrence. An average of 15 terms with the highest frequency values were selected from each document.

Similarity matrix

The inter-document similarity values for any two journal articles, d1 and d2, was calculated using the formula, s(d1, d2) = 2c/(a+b), where c is the number of common terms in the two documents, and a and b are number of terms in document d1 and document d2, respectively. The collection of all inter-document similarity values for the 100 articles formed the similarity matrix. Since 100 journal articles were considered in the study, a 100 by 100 similarity matrix was generated. The similarity ratio falls between 0 and 1. 0 indicates no similarity, while 1 indicates an equivalence of the two documents.

Clustering

A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way" (online tutorial,). A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user that must supply this criterion, in such a way that the result of the clustering will suit their needs.For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In Exclusive clustering, data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. Overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value. Hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, Probabilistic Clustering uses a completely probabilistic approach (online tutorial http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html).

Some examples of clustering algorithms include:

- K-means
- Fuzzy C-means
- Hierarchical clustering
- Mixture of Gaussians

Each of these algorithms belongs to one of the clustering types listed above. So that, K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm.

A complete hierarchical clustering algorithm is adopted to cluster the documents. Hierarchical clustering methods are of two types: Agglomerative hierarchical methods, which begin with as many clusters as objects and successively merge clusters until only one cluster remains; and Divisive hierarchical methods that begin with all objects in one cluster, and continually divide the clusters until there are as many clusters as objects. An agglomerative clustering technique was employed for this work.

Steps in Agglomerative Hierarchical Clustering

1. Start with N clusters each containing each containing a single entity, and an N x N symmetric matrix of distances (or similarities)

Let

dij = distance between item i and item j.

2. Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance). Denote the distance between these most similar clusters U and V by dUV.

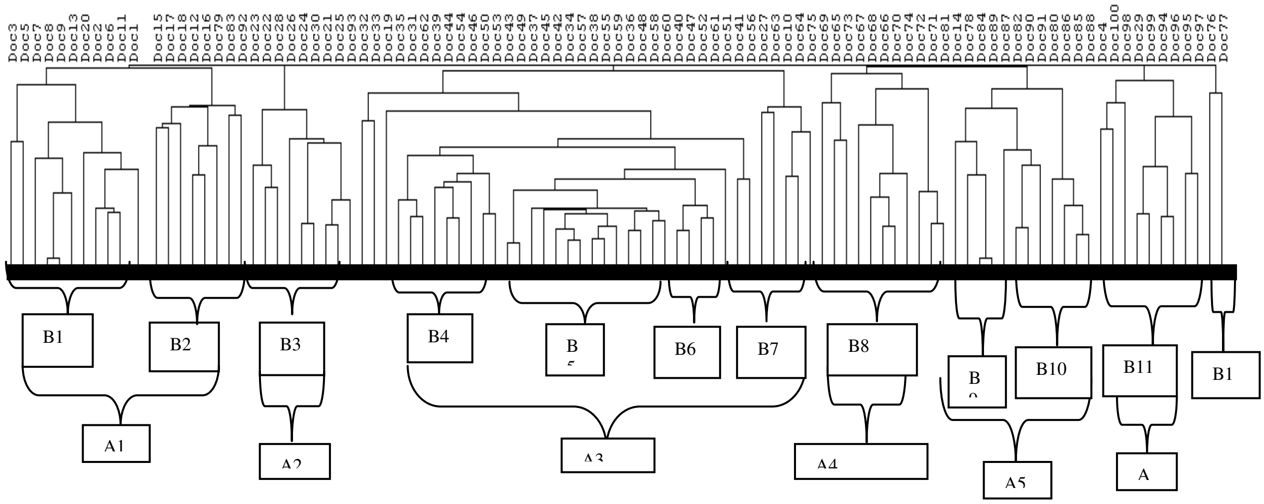
3. Merge clusters U and V into a new cluster, labeled T. Update the entries in the distance matrix by:

a. Deleting the rows and columns corresponding to clusters U and V, and

b. Adding a row and column giving the distances between the new cluster T and all the remaining clusters.

4. Repeat steps (2.) and (3.) a total of N-1 times

Figure 2: Java tree view was used to display the dendrogram of the clusters.



4. Results and Discussion

Figure 2 is a dendrogram showing the clustering pattern of the journal articles used for this study. The figure can be partitioned into seven major partitions (clusters) and twelve sub-clusters. We will first look at the sub clusters. From the figure, it is observed that the documents are clustered based on their subject area. Cluster B1 shows that docs 1, 2, 3, 5, 7, 9, 11, 13, 20 are related. Detailed study of the cluster indicated that docs 3 and 5, docs 8 and 9, docs 6 and 11 … are strongly related. These in turn are related to docs 1, 6, 13 and 20. Study of the relationship revealed that docs 3 and 5 are journal articles that dwelled on agricultural and forestry related subjects. This accounted for their having the terms “Agricultur”, “forestry” and “rangers” in common. Similarly, docs 8 and 9 are journals articles on agriculture which, centered on “effect of water harvesting methods, nitrogen – phosphorus fertilizer on plants parts”. This also accounted for their having terms like, “agricultur”, “bajopa”, “bunch”, “crop”, “fertile”, “fruit”, “harvest”, “leaf”, “palm”, “plant”, water in common. The similarity between docs 6 and 11 occurred because of the fact that both docs are journal articles related to vegetable. Meanwhile, doc 2 has a link to docs 6 and 11 at that level of clustering because the three journal articles are studies related to economic implication of the application of a farm input on crop. Furthermore, doc1 intercepted doc2, doc6, doc11 at the point of farm input – fertilizer, farm and the end product food. This is an indication that the journal articles are centered on farm products.

Cluster B2 with elements; doc15, doc17, doc18, doc12, doc16, doc79, doc83 and doc92 are purely biology related journal articles. While doc15 and doc17 are closely related due to their subject matters, doc17 and doc18 are equally correlated. Why doc15 and doc17 are journal articles on soil, doc12 and doc16 are centered on study on microbial profile on some food born microorganisms, and effect of processing on the mineral content of seed of a plant. The two journal articles centered on phytochemical and plants. Therefore, they share some common terms. Doc79 intercepted them at the point of phytochemical and journal type relation.

Furthermore, cluster B3 formed by doc22, Doc28, Doc26, Doc24, Doc30, Doc21 and Doc25 are chemistry related articles. A close look at their relation revealed that doc22 and doc28 are closely related. This is so because both journal articles dealt on subject related to composition of elements in a compound. Doc23 shared in their relationship as it dealt with the synthesis of chemical compound.

Docs 22, 23, 28, 26, 24, 30, 21, 25, and 93 formed the next sub cluster. The analysis of this cluster shows that docs 22, 23, 28, 26, 24, 30, 21, 25, and 93 are journal articles centered on aqueous chemistry.

Jumping to cluster B9, comprising doc 78, 84, 89, 87, 82, 90, 91, 80, 86, 85 and 88. It was observed that the articles are all medically centered. An analysis of their relationship showed that doc84 and 89 are closely related. Both articles treated the subject of antenatal care services. Doc78 shares some terms with them as it is equally a medical journal articles that centered on nursing career. ClusterB10 and cluster B11 comprising doc 4, 100, 98, 99, 94, 96, 95, and 97, on the other hand are bound together as they are all journal articles centered on religion. ClusterB12 shows relationship between doc76 and doc77. A close look at their relationship revealed that the two journal articles have to do with geosciences.

Looking at the major clusters, analysis indicated that the sub clusters are joined together based on their subject matter. While cluster 1 has agriculture as the main term shared by most of the docs in the cluster, some few who did not share the term agriculture are however linked to the major cluster by some agric terms. Observation showed that all journal articles in the cluster are based on agricultural related subject. Similarly, sub cluster 10 and 11 joined together because the two clusters are similar in subject area.

From the forgoing analysis of the clustering pattern, it is observed that articles that are related in their specific area formed a cluster, all specific area of applications of the field formed a cluster, all fields of the same subject area formed a cluster and all subject area of the same subject type a cluster were grouped together, all journal types are linked together. This shows that the journal articles can be classified according to their properties mentioned above. It can also be deduced that should authors and place of case study, be included in the terms, the docs would have also clustered along that line.

The observations showed clearly that journal articles have properties on which they can be arranged and searched. They can be arranged according to subject, subject area, field of study, specialized field of study, area of application, author, journal type, area of case study. Therefore, Journal article can be added to the journal database either through the subject area - the main cluster, or to the specific field of the subject area. For example, an agriculture related journal can be inserted to main cluster; cluster A1. Moving down, if the field of agriculture of the journal is plant, the article is inserted to cluster B1. Its position in cluster B1 depends on the area of plant discussed in the article. The dendrogram also presented features on which journal articles can be searched for. For example, a query such as subject = “agriculture” and field = “plant” will narrow the search to cluster B1. In the same vein, a religiously related article can be inserted through the main cluster A5 and down to the sub cluster to link to the proper field. Therefore, a vocabulary based on the properties of the journal articles can be developed so as to create a platform for organizing journal articles. This will not only enhance getting a specific result from a search on the web but will also allow the machine to understand the information on the web and make intelligent decision. This is the whole idea of semantic web, the information on the web should be resourceful not only to human but also to machine, which can only be achieved if there exist a taxonomy for set of such information on which a standard vocabulary (ontology) can be developed which will in turn be a platform to publish the information.

5. Conclusion

This study has demonstrated that journal articles have properties on which they can be arranged and searched. They can be arranged according to subject, subject area, field of study, specialized field of study, area of application, author, journal type, area of case study. Therefore, articles can be added to a collection either through the subject area - the main cluster, or to the specific field of the subject area. A vocabulary and taxonomy can thus be developed based on the properties of the journal articles so as to create a platform for organizing journal articles. In this way the goal of a semantic web, which is that all the information on the web should be meaningful, not only to humans, but also to machines, can be achieved.

The work can be extended by developing algorithms to insert new documents into the collection. The method can also be tested using much larger document collections.

REFERENCES

Berners-Lee, T. 1989. Information Management: A Proposal. CERN, March 1989, May 1990. http://www.w3.org/History/1989/proposal.html

Berners-Lee, T. The Fractal Nature of the Web, Working Draft, 1998–2005; www.w3.org/DesignIssues/Fractal.html.

Berners-Lee, T., J. Hendler, and O. Lassila 2001. The Semantic Web, Scientific American, May 2001, pp. 28-37.

Berners-Lee, T. and V. Shannon 2006. International herald 25th May

Bezdek, J.C. 1981: Pattern Recognition with Fuzzy Objective Function Algoritms. Plenum Press, New York

Hierarchical cluster analysis
http://www.clustan.com/hierarchical_cluster_analysis.html

Huang Z. and Stuckenschmidt H., Reasoning with Multi-Version Ontologies

MacQueen, J.B. 1967. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 281-297. Berkeley, University of California Press.

Moore, A.: K-means and Hierarchical Clustering. Tutorial Slides.
http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html

MusixBrainz , 2001. MusicBrainz 2.0. http://www.musicbrainz.org/MM/

Navoni,l M.???? Porter Stemmer in VISUAL BASIC 6 Algorithm Implemented as part for assignment on document visualization Brunel University

Porter, M, 1980.. An algorithm for suffix stripping, Program 14(3):130-137.

Riitta, M.K.and Miller, E. W3C Semantic Web Activity http://www.w3.org/W3C/MIT

The semantic Web An Introduction http://infomrs.net/2001/swintro/

W3C. 1999. Model and Syntax Specification. Recommendation, 22 February 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222.

W3C, 2000. Semantic Web Development. http://www.w3.org/2000/01/sw/.

Wikipaedia, Semantic Web http://en.wikipedia.org/wiki/ Semantic-Web