International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184 Volume 1, Number 1 (October 2012), pp. 45-50

http://www.meacse.org/ijcar

A Clustering Internet Search Agent for User Assistance

Ebtessam Mohamed Desouky[1] and Mahmoud Mohamed El-Khouly[2]

[1]Information Technology Department, Institute of Graduate Studies & Research

Alexandria University, Egypt

[2]Information Technology Department, Faculty of Computers & Information

Helwan University, Egypt

Abstract

The complexity of an incredibly huge, dynamic, unpredictable, and heterogeneous environment as the WWW continues to grow rapidly. This paper focuses on the design of a clustering search agent which can help users to collects web pages from search engines such as Yahoo and Google, then analyses their content and clustering it. The proposed search agent is composed of four main components: the pages retrieval module, analysis module, dimensionality reduction module and the clustering module.

Keywords: Information Retrieval, Search Agent, Web Mining.

Introduction

The WWW has become a vast resource of information. The problem is that finding information of interest is often quite difficult, due to the complexity in information organization and the quantity of information stored. The structure of the WWW is represented usually as a directed graph referred to as the Web graph. In this graph nodes represent static pages on the web, and arcs represent hyperlinks between these pages. According to Google, current Web graph contains more than 3 billion nodes and, every day; about 7.3 millions pages are added to it while many others are modified or removed, every day, more sites get created. That means these data will change fairly rapidly over short periods of time. It has become immensely important to cluster the web pages to assist search operations [1],[2].

Client-based search spiders (agents)

Research in spiders began in the early 1990's, in recent years, many client-side web spiders have been developed Agents have become a solution to assist users in their search tasks since they can retrieve, filter and organize information on behalf of their users. Broadly defined, an 'agent' is a program that can operate autonomously and accomplish unique tasks without direct human supervision (similar to human counterparts, such as real estate agents, travel agents, etc…). The basic idea of agent research is to develop software systems which engage with and assist all types of end users (clients). Such agents might act as 'spiders' on the Internet and look for relevant information, analyze meeting output on behalf of executives, or filter newsgroup articles based on induced or learned users' profiles (e.g. [3],[4],[5],[6],[7],[8]).

A brief overview of Document Clustering

Document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis. It consists of three dependent steps: preprocessing steps,

document representation and clustering algorithm. Document clustering algorithms perform several preprocessing steps including stop words removal and stemming on the documents collection. Document representation can be performed by several ways. One of the most used document representations is the vector space of weighted terms.
The classic approach attempts to adopt the well-known clustering algorithms, originally designed for numerical data, such as Hierarchical Agglomerative Clustering (HAC) or K-means, to the data of textual type. The algorithms require that for every two objects in the input collection a similarity measure be defined. The measure, which is usually calculated as a single numerical value, represents the "distance" between these objects. Objects that are "close" to each other in this sense will be placed in the same cluster. Thus, to use the classic algorithms for the new type of data, the measure of similarity between two texts must be defined. Many different approaches to this problem have been proposed [9] and [10].

Proposed Search Agent

The search agent is composed of four main components: the pages retrieval module, analysis module, dimensionality reduction module and the clustering module.
The architecture of the agent is shown in the figure (1) below.

Retrieval module

The functionality of the pages retrieval module is to explore and collect web pages from search engines. This module periodically runs as a background process. It collects web pages from search engines such as Yahoo and Google, then analyses their content [11].

Analysis module

The analysis module has three main processes. First, parse html pages' contents by specifying the web page's body and then remove all inline script tags, all undefined characters, all html spaces, outline spaces, stop words or non-informative because they will be given more weights.
Second, stemming algorithm is a process for removing the commons morphological, in flexional ending from words. This process is performed to reduce the words by stemming the plural noun to its single form and inflexed verb to its original form or root from.

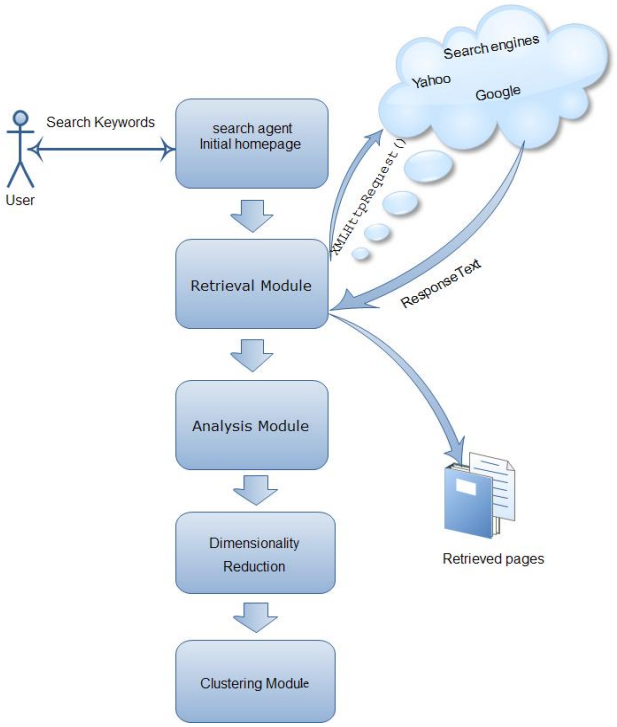

Figure 1. The architecture of Search Agent

Dimensionality Reduction Module

The output from the analysis phase is a table (array/ matrix) whose number of rows equals to the number of terms and whose number of columns equals the number of pages. The number of terms is in the order of hundreds. Since the terms correspond to features in a clustering problem, it is desirable to reduce the dimensionality of the feature space. A well-known technique to achieve this goal is the Principal Component Analysis (PCA) [12].

Clustering module

Clustering is a division of data into groups of similar objects. Each group is called a cluster and consists of objects that are similar between themselves and dissimilar to objects of other groups [13]. The idea of clustering originates from statistics applied to numerical data.

Results and discussion

In this section, an application of the clustering Internet agent is illustrated. This application involves the following steps:

1. Using the search agent to search the Web using as input the words "Intelligent" & "Agents" as keywords. The information obtained from the Yahoo search engines is preprocessed and a sample data consisting of 16 Web pages is organized into a table of 16 (rows) vs. 682 terms (columns) here is part of the table:

game	code	file	inform	artifici	intellig	comput	search	web	music
0	0	0	0	0	0	0	0	0	0
10	10	7	1	1	1	2	5	2	1
1	0	0	3	2	2	1	0	0	0
1	0	0	2	2	4	0	3	2	0
0	1	0	1	1	15	0	1	0	0
0	0	0	0	1	1	0	1	0	0
23	0	0	0	1	1	6	4	1	2
0	0	0	0	1	1	0	0	0	0
0	0	0	4	1	4	6	1	0	0
6	3	0	0	7	7	1	11	0	1
0	1	1	0	1	1	0	2	2	2
0	0	0	2	2	2	0	2	1	0
49	0	0	0	1	2	6	3	0	0
11	0	0	0	3	3	3	0	1	9
18	1	1	0	1	1	9	2	1	0
1	0	0	5	1	1	1	2	0	0

Table 1: Web pages (rows) versus terms (columns)

2-Dimensionality reduction*:* the Principal Component Analysis (PCA) is performed using Turk and Pentland algorithm [14]. The Tabular data is projected on the subspace spanned by all nonzero eigenvectors. The dimensionality of the data is reduced to16 (web page)*15 (combined features) (N=16, n=15).

3-Application of Bezdek Fuzzy-Clustering algorithm[15]: The algorithm is used iteratively to form a tree of binary classifiers for the obtained web pages. Here, the tree is of depth 2.

First Iteration*:* The initial membership values are random numbers in [0, 1] and are illustrated in figure 2. The final membership values are shown in figure 3 with a fuzzy partition coefficient f_k= 0.6014

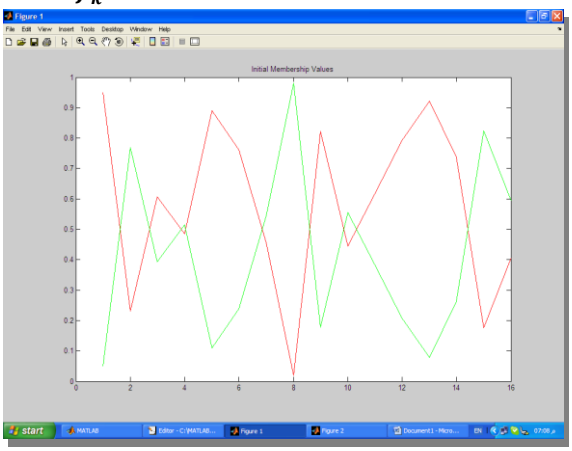

Figure 2. Initial membership values

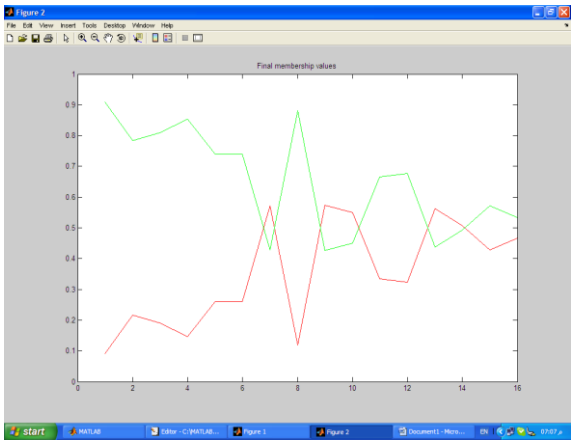

Figure 3. Final membership values

Initial Random Membership Values:

In figure 2: x-axis corresponds to web page number and y-axis corresponds to the membership value. The final membership values are shown in figure 3.

Therefore, using a membership threshold value 0.5, the clustering tree at level 1 is as shown in figure (4).

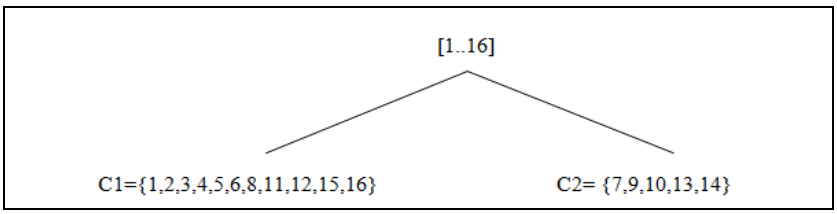

Figure 4. Clustering tree at level 1

Second Iteration: Bezdek algorithm is applied separately to the sets C1={1,2,3,4,5,6,8,11,12,15,16} (of size 11*15) and C2 = {7,9,10,13,14} (of size 5*15) to further subdivide each of them into k=2 clusters.
The initial and final membership values when the algorithm is applied to the set C1 are shown in figures (5,6) with the final fuzzy partition coefficientf_k= 0.5520
Using a membership threshold of not less than 0.53, the set C1 ={1,2,3,4,5,6,8,11,12,15,16} is subdivided into 3 groups C11={1,2,3,4,5,6,8}, C12={11,16} C13={12,15}.

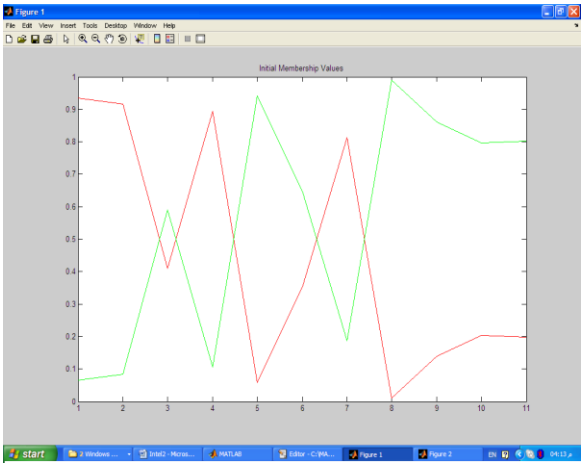

Figure 5. Initial membership values for

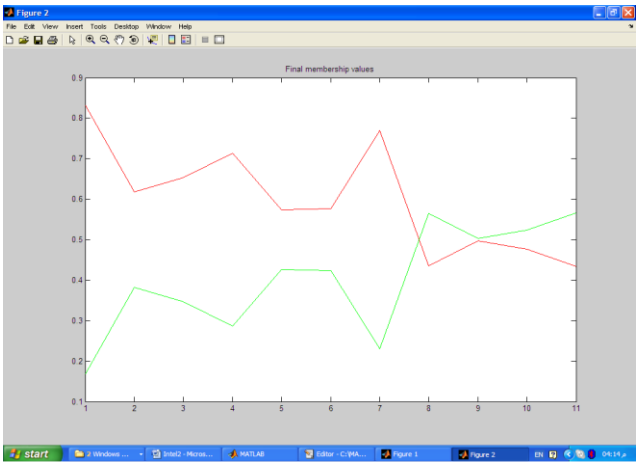

Figure 6. Final membership values for set C1

Similarly, the initial and final membership values when the algorithm is applied to the set C2 are shown in figures (7,8) with the a final fuzzy partition coefficient f_k== 0.6064.

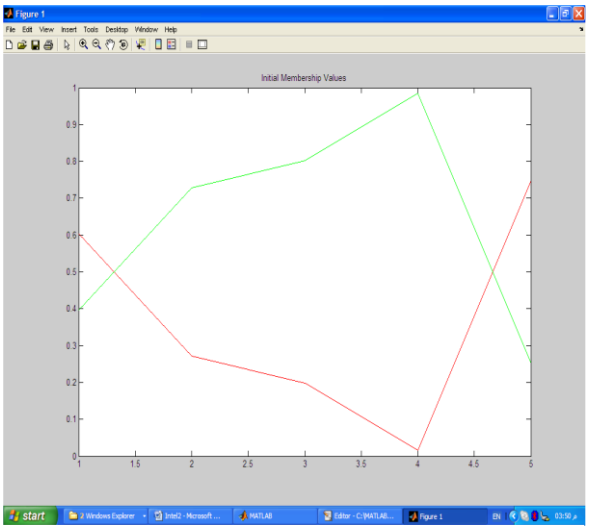

Figure 7. Initial membership values for

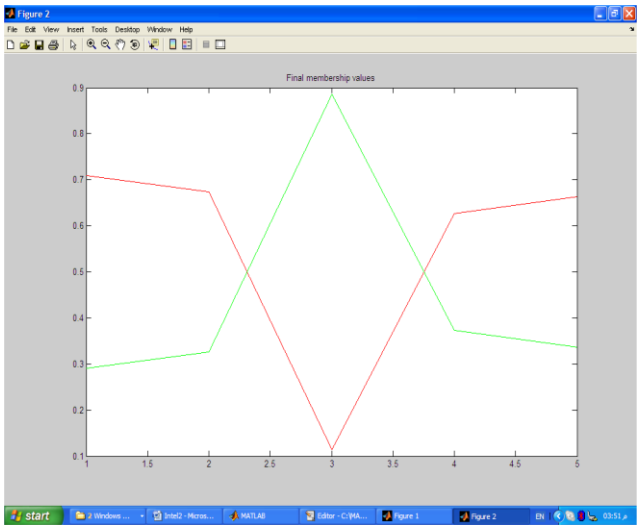

Figure 8. Final membership values for set C2

Using a membership threshold of not less than 0.53, the set C2 = {7,9,10,13,14} is subdivided into C21={7,9,13,14}and C22={10}. The final classification tree shown in figure (9) below:

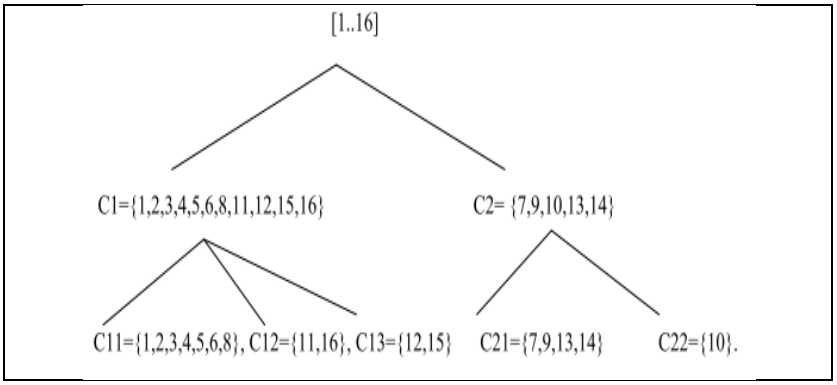

Figure 9. Final classification tree

So from figure (9) we can conclude that we can divide the 16 pages into 5 groups, pages in each group are related to each other.

Conclusion

In this paper, we proposed an Agent-based Architecture that can assist the users in surfing the WWW by using keywords entered by the users. The web pages returned by a search engine are clustered into a hierarchy binary tree, and before the clustering step the Principal Component Analysis (PCA) is used to reduce the dimensionality of the feature space.

References

[1] Sean A. Golliher ,"Search Engine Ranking Variables and Algorithms" , semj.org volume 1, supplemental issue, August 2008

[2] G. Michael Youngblood ,Web Hunting: "Design of a Simple Intelligent Web Search Agent", ACM rossroads Student Magazine ww.acm.org/crossroads/xrds5-4/webhunting.html, Summer 1999.

[3] Chen Hsinchun, Chung Yi-Ming, Marshall Ramsey, Christopher C. Yang, "An intelligent personal spider_agent/for dynamic Intranet searching", Elsevier Science B.V. Decision Support Systems 23 ,pp: 41–58, (1998).

[4] Masoud Mohammadian, "Intelligent agents for data mining and information retrieval", Library of Congress Cataloging , book ch 2 , (2004).

[5] Daniela Godoy , Silvia Schiaffino, Analia Amandi, "Interface agents personalizing Web-based tasks", available online at ww.sciencedirect.com, Cognitive Systems Research 5, pp: 207–222, (2004)

[6] Armentano, M., Godoy, D., Amandi,A.: "Personal assistants", Int. J. Human-Computer Studies, vol. 64, pp: 27–35 , (2006).

[7] Bernard j. Jansen, tracy mullen," automated gathering of web information: an in-depth examination of agents interacting with search engines", acm transactions on internet technology, vol. 6, no. 4, pp: 442–464, november 2006.

[8] C.-C. Henry Chan, "Intelligent spider for information retrieval to support mining-based price prediction for online auctioning" ,available online at www.sciencedirect.com Expert Systems with Applications 34, pp:347–356,(2008).

[9] Bernard j. Jansen, tracy mullen," automated gathering of web information: an in-depth examination of agents interacting with search engines", acm transactions on internet technology, vol. 6, no. 4, pp: 442–464, november 2006.

[10] Baeza-Yates, R., & Ribeiro-Neto, "B.Modern Information Retrieval", ACM Press,(1999).

[11] Using Web Services in ASP.NET AJAX:
http://msdn.microsoft.com/en-us/library/bb515101.aspx (Last Visited: 19 - March - 2010)

[12] I.T. Jolliffe,"Principal Component Analysis", Springer, (2002)"

[13] R.O. Duda and P.E. Hart, "Pattern Classification and Scene Analysis", John Wiley & Sons, (1973).

[14] M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, Volume 3, Number 1, 1991.

[15] Dunn, J. ," A fuzzy relative of the iso data process and its use in detecting compact, well-separated clusters", Journal of Cybernetics, vol. 3, issue (3), pp: 32-57, (1973).