Visually Exploring Concept-Based Fuzzy Clusters in Web Search Results

Orland Hoeber and Xue-Dong Yang

University of Regina, Regina, SK S4S 0A2, Canada
{hoeber, yang}@uregina.ca

Abstract. Users of web search systems often have difficulty determining the relevance of search results to their information needs. Clustering has been suggested as a method for making this task easier. However, this introduces new challenges such as naming the clusters, selecting multiple clusters, and re-sorting the search results based on the cluster information. To address these challenges, we have developed Concept Highlighter, a tool for visually exploring concept-based fuzzy clusters in web search results. This tool automatically generates a set of concepts related to the users' queries, and performs single-pass fuzzy c-means clustering on the search results using these concepts as the cluster centroids. A visual interface is provided for interactively exploring the search results. In this paper, we describe the features of Concept Highlighter and its use in finding relevant documents within the search results through concept selection and document surrogate highlighting.

1 Introduction

Users of web search systems commonly have difficulties determining the relevance of the document surrogates that comprise web search results. While some of these difficulties can be attributed to poorly crafted queries, even when the users provide a query that adequately describes their information needs, the search results are often a mixture of documents with varying degrees of relevance. This inability of web search engines to provide highly relevant search results for users' queries can be attributed to the generality of the collection of documents being searched, the ambiguity of language, and the word mismatch problem [5].

Because most searches result in a combination of relevant and irrelevant documents to the users' information needs, the users are required to make relevance decisions on a document-by-document basis. This can be time consuming, and can result in users giving up when a large portion of the search results are irrelevant. The end result is that users of web search systems often view only one to three pages worth of search results [14, 15].

One possible method for addressing this problem is to cluster the search results such that documents that are similar to one another are grouped together [11]. In such a system, the users navigate the clusters in order to narrow down the search results and avoid clusters of irrelevant documents. In the best case scenario, the users will select the relevant clusters and view lists of document surrogates in which a large portion are relevant to the users' information needs.

One of the primary challenges in clustering is determining an adequate name or description of the clusters. If this information does not correctly describe the document surrogates contained in the cluster, the users will either choose clusters that are not relevant to their information need, or will entirely miss the clusters that contain the relevant documents. Further problems with clustering web search systems include an inability to select multiple clusters simultaneously, and a lack of sorting or re-ordering the search results.

To address these drawbacks of the clustering in web information retrieval, we have developed a tool for visually exploring concept-based clusters in web search results called Concept Highlighter. This tool makes use of a concept knowledge base [9] in order to automatically generate a set of concepts related to the users' queries. The concepts generated using the concept knowledge base are used as the centroids for a single-pass fuzzy c-means clustering algorithm [2, 11] that is applied to the search results as they are retrieved via the Google API [6]. A visual representation of the fuzzy membership scores allows the users to interactively select concepts of interest and identify how this affects the clustering of the search results.

Preliminary studies have shown that this method for providing a visual representation of the fuzzy clustering results can be very effective in allowing users to narrow down the search results. Further, since the users can interactively select and un-select concepts, as well as sort and re-sort the search results, the outcome is an exploration of the search results. This ability to explore the search results allows the user to take an active role in the evaluation of the results of their web search, and is a step towards Yao's vision for web information retrieval support systems [19].

The remainder of this paper is organized as follows: An overview of clustering in web information retrieval is provided in Section 2. In Section 3, we describe our methods for obtaining the concepts and generating fuzzy c-means clusters of the search results using these concepts. Section 4 describes the methods by which the cluster membership scores are visually represented. The process for interactively exploring the results of a web search is presented in Section 5. Conclusions and future work are provided in Section 6.

2 Background

Clustering can be defined as the unsupervised classification of data objects into groups of similar objects (called clusters) [11]. Clustering has been explored for a number of years both for browsing text collections [4, 8] and for organizing web search results [20, 21]. Hearst and Pedersen validated the cluster hypothesis, showing that relevant documents tended to be more similar to each other than non-relevant documents [8]. This research provides evidence that clustering can be used to support the users' tasks of finding relevant groups of documents from a collection of search results.

Recently, a number of publicly available web search systems have been developed that provide clusters of web search results, and allow the users to browse the clusters to narrow down the set of search results. Many of these systems use hierarchical clustering algorithms, and primarily differ in the representation and interaction with the clusters. Two such systems are Vivisimo [17] and Grokker [7].

The hierarchical clustering algorithms used by these systems partition the search results at various levels of similarity [11]. The result is a tree-like structure representing the clusters, where parent clusters contain all the objects of their children clusters.

In Vivisimo, these hierarchical clusters are represented as a tree. The nodes in the tree can be expanded and collapsed in a manner similar to file directory navigation. When a tree node is selected, the document surrogates contained within that cluster are displayed in a separate frame.

In addition to providing a tree-like navigation scheme, Grokker uses a visual representation of the hierarchical cluster structure. This visual representation uses nested circles to represent the clusters and their children, and provides the ability for the users to see the sizes of the clusters and whether they contain additional children or document surrogates. Like Vivisimo, when a cluster is selected, the document surrogates that are contained within that cluster are displayed in a separate frame.

One of the challenges in any clustering system is to provide meaningful names for the clusters. Commonly, the names are generated by choosing the most frequent terms or phrases within the cluster (ignoring very common terms such as "is", "and", "the", etc.) [21]. The ability to choose meaningful descriptions of clusters in a web search system has a direct impact on the ability for the users to correctly navigate the clusters to find relevant document surrogates. If vague or misleading names are chosen for the clusters, this can lead to users choosing clusters that are not relevant to their information needs (resulting in the evaluation of documents that are likely not relevant), or not choosing clusters that are relevant to their information needs (resulting in missing documents that are relevant).

Commonly, the documents that are relevant to a users information needs will be distributed among multiple clusters. However, these systems do not easily support the exploration of multiple clusters. While it is possible to view an intermediate cluster that contains all the document surrogates of its children clusters, viewing the union of an arbitrary set of clusters is not possible. This means that if users wish to explore multiple clusters, they must do so separately.

A final difficulty with these web search clustering techniques is that they do not provide any additional information regarding the organization of the document surrogates within the clusters. When a cluster is selected, the documents are listed in the same order as provided by the underlying search engine. There is no indication of which documents are most similar to the cluster centroid, or the degree of membership to the cluster.

To address these shortcomings of the web search clustering systems, we have developed Concept Highlighter, a tool that provides a visual and interactive interface to concept-based fuzzy clusters of web search results. In this tool, the clusters are named using the concept names; multiple clusters can be selected generating a union of the clusters; and the search results are re-sorted based on their membership score. Information visualization techniques are used to visually represent the fuzzy membership scores of the document surrogates in an abstract and compact form, allowing the users to visually process and interpret this information.

3 Fuzzy Clustering Using Concepts

The first step in generating the concept-based fuzzy clusters is to obtain a set of concepts associated with the users' queries. The source of the conceptual information is a concept knowledge base that was originally devised for query expansion [9, 10]. This concept knowledge base contains relationships between concepts and the terms have been used to describe them. The ACM Computing Classification System [1] was used as the source of the conceptual knowledge for the prototype tool, resulting in a concept knowledge base specifically for the computer science domain.

The process for obtaining the concepts that are related to the users' queries is similar to the process for generating the query space as described in [10]. The query terms are first processed using Porter's stemming algorithm [12], which removes the prefixes and suffixes from terms to generate the root words, called stems. These stems are matched to the stems in the concept knowledge base, and the nearest concepts are selected. For each of these concepts, the set of stems that are nearest to the concept are selected from the knowledge base. Each of these sets will contain one or more of the original query term stems, plus additional stems that are not present in the query.

In our previous work, the resulting query space was used to allow the users to interactively refine their queries. In this work, we instead use this query space to identify potential cluster centroids that may be relevant to the users' information needs. For each concept, a vector is created using the set of stems that were selected from the concept knowledge base. The weight of the link between the concept-stem pair is used to set the magnitude of the concept vector in the dimension associated with the stem.

Therefore, as a result of this query space generation, a set of concept vectors $C = \{c_1, c_2, \ldots, c_m\}$ are generated. If the total number of unique stems that were selected from the concept knowledge base is p, then the dimension of all vectors c_i $(i = 1 \ldots m)$ is p. Further, the magnitude of the vector c_i $(i = 1 \ldots m)$ on dimension j $(j = 1 \ldots p)$ is given by the concept knowledge base weight between concept i and term j.

After the concepts have been obtained from the concept knowledge base, and the concept vectors have been created, the users' queries are sent to the Google API [6]. As each of the document surrogates are retrieved, a single-pass fuzzy c-means clustering algorithm [2, 11] is performed. The title and snippet from the document surrogate are processed using Porter's stemming algorithm [12], and the frequency of each unique stem is calculated. These frequencies are used to generate vectors for each of the document surrogates. Although some argue against using term frequencies (TF) as the sole source of information in a text retrieval system [13], using other global information such as the inverse document frequency (IDF) is not feasible when the document surrogate vectors need to be generated as each document surrogate is retrieved (to achieve a near real-time web information retrieval system).

Given a set of concept vectors $C = \{c_1, c_2, \ldots, c_m\}$ and a document surrogate vector d_i, the fuzzy membership of document surrogate d_i with respect to concept c_j is given by:

$$u_{i,j} = \frac{1}{\sum_{k=1}^{m} (\frac{sim(d_i,c_j)}{sim(d_i,c_k)})^2}$$

In this calculation, the similarity between a document surrogate vector and a concept vector is given by the Euclidean distance metric [11]:

$$sim(x_i, x_j) = (\sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2)^{1/2}$$

Normally, when evaluating the document surrogates, all unique stems would contribute to the construction of the document surrogate vector. However, since the distance calculations in this single-pass fuzzy clustering algorithm are always between concept vectors and document surrogate vectors, we only need to consider the stems that are already present in the concept vectors. This reduction in the dimension of the document surrogate vectors results in an increase in the speed at which the fuzzy clusters are generated. In our prototype system, the fuzzy cluster membership scores are calculated as quickly as the underlying search engine can provide the document surrogates to the system.

While it is common to run the fuzzy c-means clustering algorithm in multiple passes, each time re-calculating the centroids of the clusters, we only run the algorithm in a single pass resulting in a fuzzy membership score for each concept-document surrogate pair. This ensures that the fuzzy clusters remain centred around the concepts.

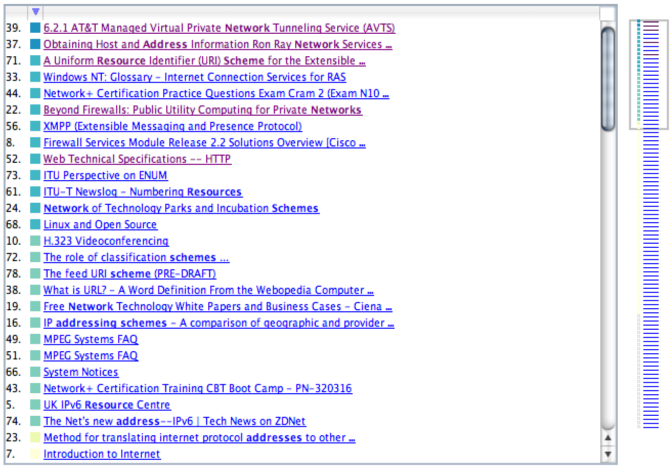
Since the concepts represent the centroids of the clusters, the clusters can be named using the concept names. This is a valuable benefit since the names of the concepts are derived from the source knowledge upon which the concept knowledge base was constructed (in this case, the ACM Computing Classification System). Further, since the clusters always remain centred on the concepts, they are independent of the search results. Therefore, while two similar queries will result in two different sets of search results, they will commonly result in a very similar set of concepts. This can be beneficial as the users learn which concepts are of interest to their general information seeking needs (and can be extended in the future to support personalized concept selection).

4 Visual Representation of Membership Scores

Information visualization takes advantage of the human visual information processing systems by generating graphical representations of data or concepts [18]. The cognitive activity involved in viewing and processing a visual representation allows users to gain understanding or insight into the underlying data. With respect to the visualization of fuzzy clusters, the ultimate goal is to allow users to *see* the clusters without limiting their ability to view the entire set of document surrogates.

Concept Highlighter provides a compact list-based representation at two levels of detail: the overview map shows the membership scores for the first 100 documents returned by the Google API in a single compact list; the detail view shows approximately 25 document surrogates at a time. A screenshot of these two levels of detail are shown in Figure 1.

Our preliminary studies have shown that most users have a preference for a compact representation of web search results, which can more easily be visually scanned. As such, the only persistent information from the document surrogate provided in the detail view is the title. The snippet and URL associated with each document surrogate can be accessed as needed via a tool tip. Additionally, the detail view provides the document surrogate number, allowing the user to easily identify the degree of importance placed on this document surrogate by the underlying search engine algorithms.



39. 6.2.1 AT&T Managed Virtual Private **Network** Tunneling Service (AVTS)
37. Obtaining Host and **Address** Information Ron Ray **Network** Services ...
71. A Uniform **Resource** Identifier (URI) **Scheme** for the Extensible ...
33. Windows NT: Glossary – Internet Connection Services for RAS
44. Network+ Certification Practice Questions Exam Cram 2 (Exam N10 ...
22. Beyond Firewalls: Public Utility Computing for Private **Networks**
56. XMPP (Extensible Messaging and Presence Protocol)
8. Firewall Services Module Release 2.2 Solutions Overview [Cisco ...
52. Web Technical Specifications -- HTTP
73. ITU Perspective on ENUM
61. ITU-T Newslog – Numbering **Resources**
24. **Network** of Technology Parks and Incubation **Schemes**
68. Linux and Open Source
10. H.323 Videoconferencing
72. The role of classification **schemes** ...
78. The feed URI **scheme** (PRE-DRAFT)
38. What is URL? – A Word Definition From the Webopedia Computer ...
19. Free **Network** Technology White Papers and Business Cases – Ciena ...
16. IP **addressing schemes** – A comparison of geographic and provider ...
49. MPEG Systems FAQ
51. MPEG Systems FAQ
66. System Notices
43. Network+ Certification Training CBT Boot Camp – PN-320316
5. UK IPv6 **Resource** Centre
74. The Net's new **address**--IPv6 | Tech News on ZDNet
23. Method for translating internet protocol **addresses** to other ...
7. Introduction to Internet

Fig. 1. The visual representation of the web search results consists of an overview map (right) and a detail window (left). These search results were returned from the query "addressing schemes resources networks", and show the fuzzy membership score when the concepts "computer-communication networks: network architecture and design" and "operating systems: communications management" are enabled. The document surrogates with the purple links are those that have been viewed by the user in this search session.

Since the spatial position of an object and its colour can be perceptually separated, colour coding of the fuzzy membership scores can be used without interfering with the spatial layout of the data [18]. In many cases, colour can be pre-attentively processed, allowing the information to be absorbed by the users faster than if they were required to read the corresponding numerical values [18]. While identifying specific values in the colour scale used in Concept Highlighter may not be pre-attentively processed, identifying a relative ordering as well as a few high values from many low values will be processed faster than reading the numerical values.

The choice of a colour scale is not as simple as it might seem. Since we need to represent an ordered sequence of values, a colour sequence that varies monotonically on at least one colour channel is required [16, 18]. A set of nine perceptually distinct colours on a yellow-green-blue colour scale were chosen to represent the fuzzy membership scores. This colour scale varies on all three colour channels: luminance, yellow-blue, and red-green. The ColorBrewer application [3] was used to select this colour scale.

In order to allow the users to remain aware of the location of the detail view with respect to the larger set of documents represented in the overview map, a grey box is used to indicate the correspondence between these two coordinated views. Together, these views allow the user to both investigate the document surrogates in detail, as well as gain insight into the features of the entire set of search results displayed.

5 Interactive Search Results Exploration

Users of Concept Highlighter can interactively explore the search results in a number of different ways. A list of the concepts matched to the users' queries is provided at the top of the display. Beside each concept is a checkbox which can be used to enable or disable the corresponding fuzzy cluster.

When the user checks a cluster, the fuzzy membership scores for all the document surrogates is visually represented in both the overview map and detail view. The user may check multiple concepts, the result of which generates a summation of the fuzzy membership scores corresponding to the selected concepts. Therefore, as multiple concepts are selected, the document surrogates that are nearer to both clusters are represented with a darker colour on the colour scale, indicating their higher fuzzy membership score.

As the documents that belong to the selected clusters are highlighted, the user may visually inspect both the overview map and the detail view to find relevant documents. Clicking on any location in the overview map will automatically scroll the detail view to that location. Therefore, the users can easily scan the entire 100 documents shown in the overview map, and jump to locations of interest based on the fuzzy membership score visualization.

To make it easier for the users to systematically view the set of documents that are contained within the selected fuzzy clusters, a sorting mechanism is supported in the detail view, and is enabled by default. Clicking the column header above the colour codes for the fuzzy membership scores will disable the sort. Any changes to the sorting will be instantly reflected in the overview map as well detail view.

The interactive nature of concept cluster selection, and the sorting of the documents based on the total fuzzy membership score allows the users to interactively explore the search results. Using the fuzzy clusters as a means for organizing the search results in this exploration process can help in bringing documents that are relevant to the users' information needs into focus, even if these documents are deep in the search results.

An example of a scenario in which a user performs a search for "addressing schemes resources networks" is provided in Figure 2. A video showing this scenario is provided on the author's web site [1].

6 Conclusions & Future Work

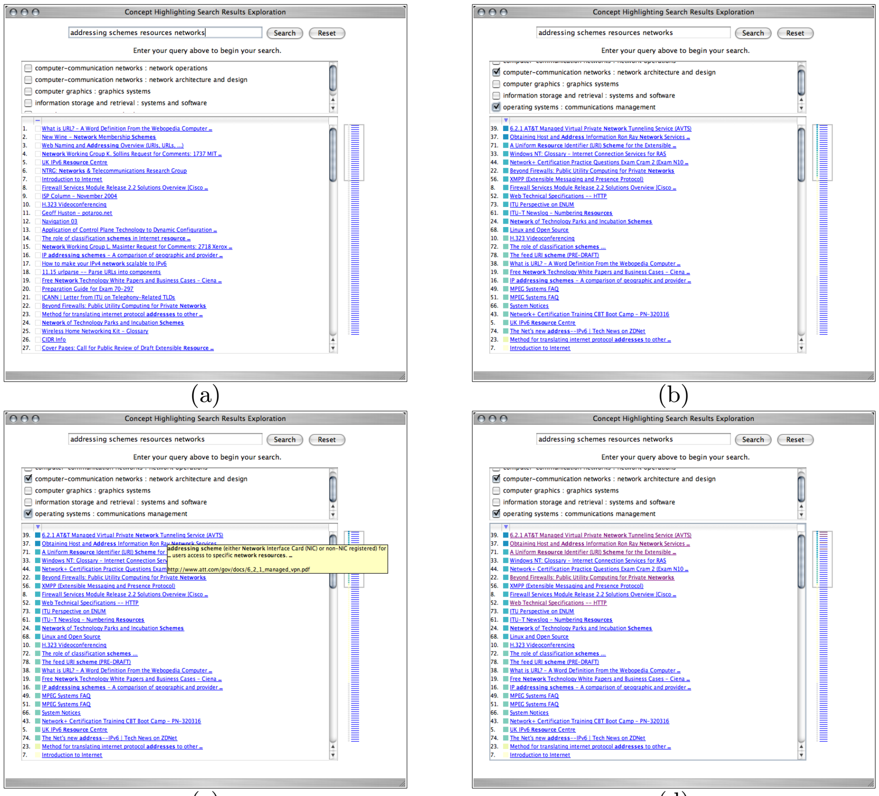
Even for well crafted queries, the results of web searches often contain document surrogates of varying degrees of relevance to the users' information seeking goals. Clustering of the search results allows the users to navigate the clusters in order to narrow down the set of search results to a smaller collection containing a larger ratio of relevant documents. However, most web search clustering systems use simple keyword-based cluster naming techniques; do not allow the users to select multiple clusters simultaneously; and do not organize the search results once a cluster is selected.

In this paper, we described Concept Highlighter, a tool for generating concept-based fuzzy clusters of web search results, and an interface for visually representing the fuzzy membership scores and interactively exploring web search results. The visual exploration of the concept-based fuzzy clusters allows the users to interactively select the concepts they think may be relevant to their information seeking goal, and see the results of these concept selections in the highlighting of the document surrogates that belong to the corresponding fuzzy clusters.

The ability of Concept Highlighter to allow the users to find relevant document surrogates depends on the ability of the tool to match the users' queries to the concept knowledge base. Further, if there are few concepts returned, or if all the concepts returned are relevant to the users' information needs, the ability to assist the users in narrowing down the search results is diminished. This can occur when the users' queries are very specific, and will often lead to very specific search results.

More often, the users' queries are less specific. This results in multiple concepts being selected from the concept knowledge base, and a more general collection of documents being returned from the search engine. It is in these situations Concept Highlighter can assist the users in finding relevant documents. The interactive exploration of the web search results using the concept-based fuzzy clusters can lead the users to groups of document surrogates that are relevant, and away from groups of document surrogates that are less relevant.

[1] `http://www.cs.uregina.ca/~hoeber/ConceptHighlighter/`



addressing schemes resources networks

Enter your query above to begin your search.

□ computer–communication networks : network operations
□ computer–communication networks : network architecture and design
□ computer graphics : graphics systems
□ information storage and retrieval : systems and software

1. What is URL? – A Word Definition From the Webopedia Computer ...
2. New Wine – **Network** Membership **Schemes**
3. Web Naming and **Addressing** Overview (URIs, URLs, ...)
4. **Network** Working Group K. Sollins Request for Comments: 1737 MIT ...
5. UK IPv6 **Resource** Centre
6. NTRG: **Networks** & Telecommunications Research Group
7. Introduction to Internet
8. Firewall Services Module Release 2.2 Solutions Overview [Cisco ...
9. ISP Column – November 2004
10. H.323 Videoconferencing
11. Geoff Huston – potaroo.net
12. Navigation 03
13. Application of Control Plane Technology to Dynamic Configuration ...
14. The role of classification **schemes** in Internet **resource** ...
15. **Network** Working Group L. Masinter Request for Comments: 2718 Xerox ...
16. IP **addressing schemes** – A comparison of geographic and provider ...
17. How to make your IPv4 **network** scalable to IPv6
18. 11.15 urlparse -- Parse URLs into components
19. Free **Network** Technology White Papers and Business Cases – Ciena ...
20. Preparation Guide for Exam 70–297
21. ICANN | Letter from ITU on Telephony–Related TLDs
22. Beyond Firewalls: Public Utility Computing for Private **Networks**
23. Method for translating internet protocol **addresses** to other ...
24. **Network** of Technology Parks and Incubation **Schemes**
25. Wireless Home Networking Kit – Glossary
26. CIDR Info
27. Cover Pages: Call for Public Review of Draft Extensible **Resource** ...

(a)

addressing schemes resources networks

Enter your query above to begin your search.

☑ computer–communication networks : network architecture and design
□ computer graphics : graphics systems
□ information storage and retrieval : systems and software
☑ operating systems : communications management

39. 6.2.1 AT&T Managed Virtual Private **Network** Tunneling Service (AVTS)
37. Obtaining Host and **Address** Information Ron Ray **Network** Services ...
71. A Uniform **Resource** Identifier (URI) **Scheme** for the Extensible ...
33. Windows NT: Glossary – Internet Connection Services for RAS
44. Network+ Certification Practice Questions Exam Cram 2 (Exam N10 ...
22. Beyond Firewalls: Public Utility Computing for Private **Networks**
56. XMPP (Extensible Messaging and Presence Protocol)
8. Firewall Services Module Release 2.2 Solutions Overview [Cisco ...
52. Web Technical Specifications -- HTTP
73. ITU Perspective on ENUM
61. ITU–T Newslog – Numbering **Resources**
24. **Network** of Technology Parks and Incubation **Schemes**
68. Linux and Open Source
10. H.323 Videoconferencing
72. The role of classification **schemes** ...
78. The feed URI **scheme** (PRE–DRAFT)
38. What is URL? – A Word Definition From the Webopedia Computer ...
19. Free **Network** Technology White Papers and Business Cases – Ciena ...
16. IP **addressing schemes** – A comparison of geographic and provider ...
49. MPEG Systems FAQ
51. MPEG Systems FAQ
66. System Notices
43. Network+ Certification Training CBT Boot Camp – PN–320316
5. UK IPv6 **Resource** Centre
74. The Net's new **address**--IPv6 | Tech News on ZDNet
23. Method for translating internet protocol **addresses** to other ...
7. Introduction to Internet

(b)

addressing schemes resources networks

Enter your query above to begin your search.

☑ computer–communication networks : network architecture and design
□ computer graphics : graphics systems
□ information storage and retrieval : systems and software
☑ operating systems : communications management

39. 6.2.1 AT&T Managed Virtual Private **Network** Tunneling Service (AVTS)
37. Obtaining Host and **Address** Information Ron Ray **Network** Services
71. A Uniform **Resource** Identifier (URI) **Scheme** for
33. Windows NT: Glossary – Internet Connection Serv
44. Network+ Certification Practice Questions Exam
22. Beyond Firewalls: Public Utility Computing for Private **Networks**
56. XMPP (Extensible Messaging and Presence Protocol)
8. Firewall Services Module Release 2.2 Solutions Overview [Cisco ...
52. Web Technical Specifications -- HTTP
73. ITU Perspective on ENUM
61. ITU–T Newslog – Numbering **Resources**
24. **Network** of Technology Parks and Incubation **Schemes**
68. Linux and Open Source
10. H.323 Videoconferencing
72. The role of classification **schemes** ...
78. The feed URI **scheme** (PRE–DRAFT)
38. What is URL? – A Word Definition From the Webopedia Computer ...
19. Free **Network** Technology White Papers and Business Cases – Ciena ...
16. IP **addressing schemes** – A comparison of geographic and provider ...
49. MPEG Systems FAQ
51. MPEG Systems FAQ
66. System Notices
43. Network+ Certification Training CBT Boot Camp – PN–320316
5. UK IPv6 **Resource** Centre
74. The Net's new **address**--IPv6 | Tech News on ZDNet
23. Method for translating internet protocol **addresses** to other ...
7. Introduction to Internet

addressing scheme (either **Network** Interface Card (NIC) or non–NIC registered) for ... users access to specific **network resources**. ...

http://www.att.com/gov/docs/6_2_1_managed_vpn.pdf

(c)

addressing schemes resources networks

Enter your query above to begin your search.

☑ computer–communication networks : network architecture and design
□ computer graphics : graphics systems
□ information storage and retrieval : systems and software
☑ operating systems : communications management

39. 6.2.1 AT&T Managed Virtual Private **Network** Tunneling Service (AVTS)
37. Obtaining Host and **Address** Information Ron Ray **Network** Services ...
71. A Uniform **Resource** Identifier (URI) **Scheme** for the Extensible ...
33. Windows NT: Glossary – Internet Connection Services for RAS
44. Network+ Certification Practice Questions Exam Cram 2 (Exam N10 ...
22. Beyond Firewalls: Public Utility Computing for Private **Networks**
56. XMPP (Extensible Messaging and Presence Protocol)
8. Firewall Services Module Release 2.2 Solutions Overview [Cisco ...
52. Web Technical Specifications -- HTTP
73. ITU Perspective on ENUM
61. ITU–T Newslog – Numbering **Resources**
24. **Network** of Technology Parks and Incubation **Schemes**
68. Linux and Open Source
10. H.323 Videoconferencing
72. The role of classification **schemes** ...
78. The feed URI **scheme** (PRE–DRAFT)
38. What is URL? – A Word Definition From the Webopedia Computer ...
19. Free **Network** Technology White Papers and Business Cases – Ciena ...
16. IP **addressing schemes** – A comparison of geographic and provider ...
49. MPEG Systems FAQ
51. MPEG Systems FAQ
66. System Notices
43. Network+ Certification Training CBT Boot Camp – PN–320316
5. UK IPv6 **Resource** Centre
74. The Net's new **address**--IPv6 | Tech News on ZDNet
23. Method for translating internet protocol **addresses** to other ...
7. Introduction to Internet

(d)

Fig. 2. A common usage scenario would begin with the user entering their query and viewing the search results (a). The user may then check the concepts that are relevant to their query which sorts the document surrogates based on their fuzzy membership score (b). The users may view the snippet and URL contained in the tool tip (c). The link colour changes as documents are viewed, allowing the users to easily identify what they have previously seen (d).

Preliminary investigations have shown this tool to be quite effective in bringing relevant documents to the users' attention; a more systematic study is currently underway to determine the benefits of this work over other clustering methods and simple list-based representations. Since the concept knowledge base used in this work is specific to the computer science domain, the usefulness for general web searching is somewhat limited. The development of a more general concept knowledge base would broaden the applicability of this tool to more general web searching. Other future work includes the integration of this tool with our larger research project of developing a complete framework for a visual and interactive web information retrieval support system.

References

1. ACM. ACM computing classification system. http://www.acm.org/class/.
2. James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
3. Cynthia A. Brewer. www.colorbrewer.org, 2005.
4. Douglass Cutting, David Karger, Jan Pedersend, and John Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.
5. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 1987.
6. Google. Google web API. www.google.com/apis/, 2005.
7. Grokker. http://www.grokker.com/.
8. Marti Hearst and Jan Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996.
9. Orland Hoeber, Xue-Dong Yang, and Yiyu Yao. Conceptual query expansion. In *Proceedings of the Atlantic Web Intelligence Conference*, 2005.
10. Orland Hoeber, Xue-Dong Yang, and Yiyu Yao. Visualization support for interactive query refinement. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 2005.
11. A.K. Jain, M.N.Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), September 1999.
12. Martin Porter. An algorithm for suffix stripping. *Program*, 14(3), 1980.
13. S. E. Robertson and K. Sparck Jones. Simple proven approaches to text retrieval. Technical Report TR356, Cambridge University Computer Laboratory, 1997.
14. Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1), 1999.
15. Amanda Spink, Dietmar Wolfram, B. J. Jansen, and Tefko Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 2001.
16. Edward Tufte. *Envisioning Information*. Graphics Press, 1990.
17. Vivisimo. http://www.vivisimo.com/.
18. Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2004.
19. Yiyu Yao. Information retrieval support systems. In *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, 2002.
20. Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.
21. Oren Zamir and Oren Etzioni. Grouper: A dynamic clustering interface to web search results. In *Proceedings of the Eighth International World Wide Web Conference*, 1999.