The ATHENS System for Novel Information Discovery

N. Vats D.B. Skillicorn
School of Computing
Queen's University
{vats,skill}@cs.queensu.ca

October 2004
External Technical Report
ISSN-0836-0227-
2004-489

Department of Computing and Information Science
Queen's University
Kingston, Ontario, Canada K7L 3N6

Document prepared October 13, 2004

Abstract

Discovering novel information, that is information whose existence is not suspected or for which suitable descriptors are not known, is difficult in large information repositories. The ATHENS system finds novel information, given an initial set of keywords representing known information, using a focused and contextualized search. We present the design of the system and show its use in a number of different settings.

The Athens System for Novel Information Discovery

N. Vats and D.B. Skillicorn
{vats,skill}@cs.queensu.ca

Abstract: Discovering novel information, that is information whose existence is not suspected or for which suitable descriptors are not known, is difficult in large information repositories. The Athens system finds novel information, given an initial set of keywords representing known information, using a focused and contextualized search. We present the design of the system and show its use in a number of different settings.

1 The Problem of Novel Information Discovery

The world is full of information repositories, the largest of which is the web. Using these repositories is difficult because it is hard to find particular information (repositories are often extremely large) and because it is hard to assess the quality of information, once found. These problems were not so serious in earlier information repositories (e.g. libraries) because human gatekeepers decided both what should be included and where it should fit.

Search engines and other information retrieval tools play a major role in making information repositories useful because they reduce the effective size (by returning only a minute fraction of the available information), and by ordering the information according to some *measure* of its significance. For example, Google uses sophisticated algorithms to decide the order in which retrieved pages should be presented; systems such as Webrat [8] organize the retrieved pages into clusters by topic, and so on.

Information retrieval tools are effective when appropriate descriptors for information, such as search terms, are known. However, they do not help with the problem of discovering information whose existence is not known, or for which it is not clear how to formulate descriptors.

In earlier (and smaller) information repositories, there were two ways to discover such novel information. The first was serendipitous encounter, in which novel information might be encountered, in passing, while browsing. The existence of search engine toolbars in browsers is rapidly supplanting browsing, to the extent that even the use of bookmarks is diminishing in favor of search. Hence it is less and less likely that a user will encounter useful novel information 'by accident'. The second way to discover novel information was to use human-maintained hubs such as Yahoo. As the amount of information available on the web grows, these hubs are becoming patchy as the effort required to maintain them becomes too large to be economically viable.

Using information retrieval tools directly for novel information discovery quickly founders because of the rich interconnection structure of typical documents. A simple example illustrates this explosion problem in the web. A search at Google with the query *web browsing* retrieves 3,170,000 web pages. However, the first page that contains information about *wikis* is ranked 139th in the list of pages. A user has to be determined even to discover the existence of wikis. A further search using the query *wiki* produces a further 12,200,000 pages. Hence, although wikis are an important technology related to web browsing, the connection is hard to find, and their mutual connection to other relevant topics even harder.

In this paper we address the problem of finding novel information in information repositories, that is how a user discovers something of whose existence he or she is not already aware. We present the Athens system, a tool for discovering novel information in large information repositories.

The current version works for the web, and is piggybacked onto the Google search engine, but these limitations are not fundamental. A user gives ATHENS a set of keywords representing known information; the system returns information that is relevant to the initial information but indirectly connected to it (the assumption is that the user will already know, or can easily find out, relevant and directly-connected information). In other words, ATHENS suggests what a user, who already knows all about the information implicit in the initial keywords, should find out about next.

As well as being used as a novel information discovery tool, ATHENS can also be used as:

- A *learning* tool, pointing out what is appropriate to learn next;
- A *backgrounding* tool, discovering background information related to a topic;
- An *intelligence* tool, since indirect links can sometimes indicate concealed direct links;
- A *planning* tool, suggesting areas in which an individual or organization is well-positioned to work.

2 The Prototype Athens System

2.1 Novel Information

We propose the following definition to encapsulate the idea of novel information:
Novel Information, *relative to a set of terms, is relevant and useful information that is not present in the set of pages returned by searching on any of the terms.*

In terms of the similarity connections in the web or other information repositories, novel information is connected to the initial keywords, and so is relevant, but is not connected directly and so is not easily found using search engines.

We can infer the following characteristics about the nature of novel information. First, novel information is indirectly connected to current knowledge. However, the web is a small-world graph, so almost every page is connected to every other page by a short path. In order to be interesting and relevant, novel information should therefore be connected by *multiple* indirect paths to current knowledge.

Second, novel information should be at least two steps away along such paths, since information that is one step away is easily discovered by search engines and is likely to be already known to a user.

These characteristics suggest a strategy for discovering novel information. Clearly, a breadth-first search from the documents described by the initial terms is doomed, because of the explosion in the number of pages at each distance. What is needed is to focus the search along paths that are, in some sense, the most interesting. ATHENS implements this strategy.

2.2 System Design

2.2.1 Universe of the System

The prototype ATHENS system has been developed for discovering novel information in the web. However, the design principles involved represent a generic framework that can be applied to any information repository of documents, provided the following conditions are satisfied:

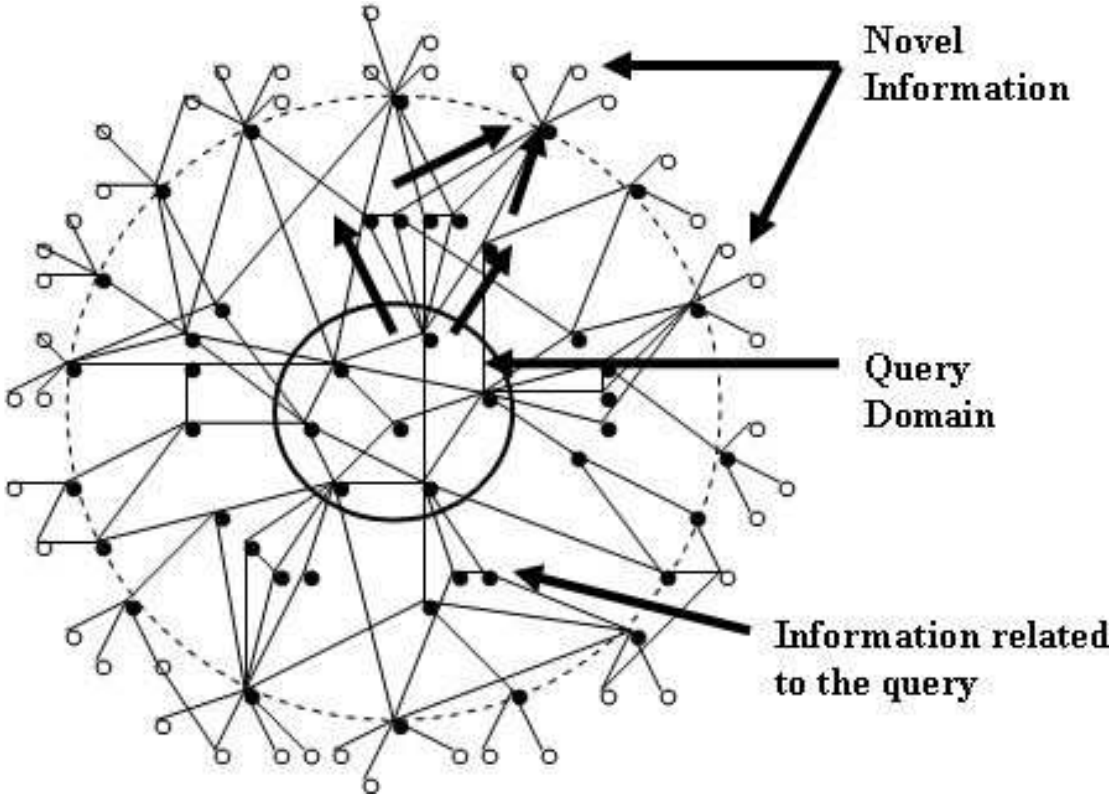

Figure 1: A Simplistic Illustration of the Similarity Graph of the Web

- **Search Engine**: A search interface must be available to retrieve sets of documents in response to a query, and it must order them by some quality metric.
- **Language Independence**: A linguistic component must be available for tagging document text so that particular parts of speech can be extracted.
- **Language Sample**: An ordered list of words in the language must be available, together with word frequencies. This is required to determine which parts of the content of documents is most interesting. A stop word list for the language is also required to eliminate irrelevant words (e.g. markup tag names).

ATHENS can be applied to the web, using any restrictions made possible by an underlying search engine (for example, restricting the search to certain kinds of documents, certain languages, certain domains, and certain time periods). It can also be applied to any other information repository for which the conditions above can be satisfied.

2.2.2 Spectral Graph Partitioning

Most information retrieval systems use a vector representation for documents. Similarity between documents, or between a set of search terms and documents, is represented by the cosine of the angle between the appropriate vectors. Similarity may be considered in a direct vector space representation, or in a lower-dimensional space using singular value decomposition (or LSI [7]).

We use this framework for classifying and clustering the documents returned by the underlying search mechanism, but we use spectral graph partitioning as the fundamental clustering technique. Spectral graph partitioning [5] is based on the eigenvalues and eigenvectors of the Laplacian matrix of a graph. The goal is to deduce the principal properties and structure of a graph from its graph spectrum.

In a graph G, let d_v denote the degree of a vertex v. If A is an adjacency matrix of a graph, then we consider a matrix l defined as follows:

$$l(u,v) = \begin{cases} d_v & \textit{if } u = v \\ -1 & \textit{if } u \textit{ and } v \textit{ are adjacent} \\ 0 & \textit{otherwise} \end{cases} \tag{1}$$

Let T denote the diagonal matrix with the (v,v)-th entry having value d_v. The Laplacian[1] of G is defined to be the matrix:

$$L(u,v) = \begin{cases} 1 & \textit{if } u = v \\ -\frac{w(u,v)}{\sqrt{d_u d_v}} & \textit{if } u \textit{ and } v \textit{ are adjacent} \\ 0 & \textit{otherwise} \end{cases} \tag{2}$$

We can write

$$L(u,v) = T^{-1/2}\ l\ T^{-1/2} \tag{3}$$

The Laplacian matrix has the characteristic that the eigenvector associated with the second smallest eigenvalue is used to define a semi-optimal partition graph. This eigenvector is known as the *Fiedler Vector* and a partition using this vector gives a guaranteed approximation to the optimum solution.

This spectral analysis can be extended to get a multi-way partitioning by selecting more than one eigenvector at a time. Algorithms have been proposed that use k vectors simultaneously to give k-way partitioning of the graph.

2.2.3 Word Frequency Issues

Zipf's Law [16] is an empirical observation that the frequency of words in natural languages decays as a power function of their rank. In other words, when words in sample texts are arranged in decreasing frequency order and a rank ordering is assigned (that is, rank 1 is assigned to the most frequent word, rank 2 to the next most frequent word and so on), then the frequency of occurrence, $f(i)$, of the i^{th} word in the frequency order is given by the function

$$f(i) \sim \frac{1}{i^a} \tag{4}$$

with the exponent a close to unity.

The best illustration of Zipf's law comes from studying the statistical properties of English words in the British National Corpus or BNC [3]. The BNC is a huge sample consisting of more than 100 million words in written and spoken English, along with their frequencies. Figure 2 plots the frequency of the 300 most frequent words in the entire BNC sample against their rank. Clearly the frequency decays as a power function of the rank and hence obeys Zipf's Law. Further the log–log frequency distribution (see Figure 3) is a straight line, so Zipf's law is obeyed.

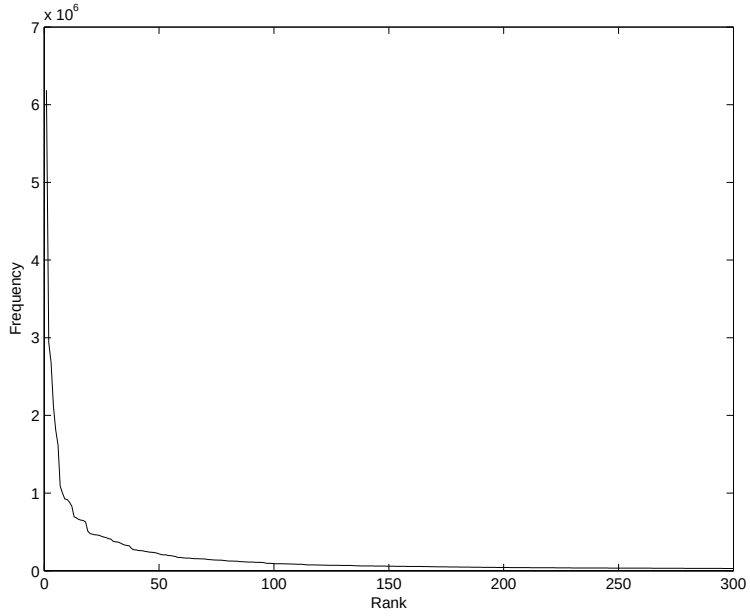

Figure 2: Zipf Distribution of English Words in the BNC Corpus

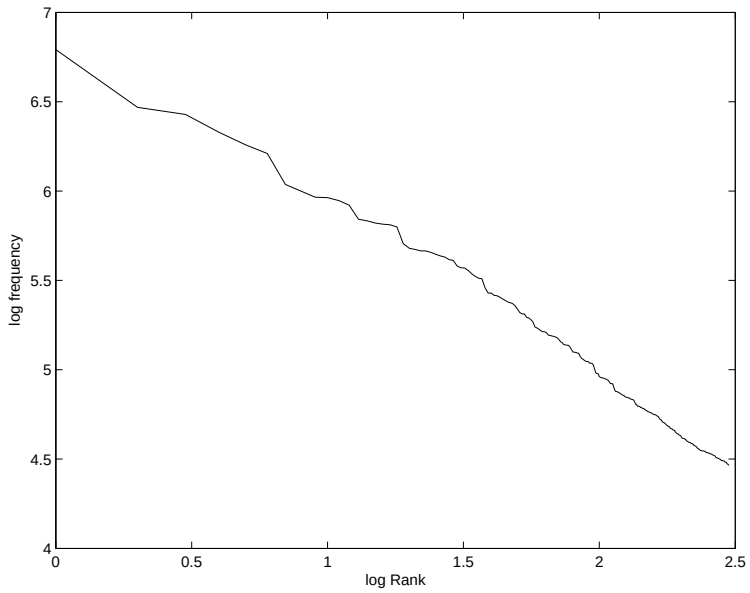

Figure 3: Zipf Distribution: log–log Plot

Significant entities and concepts are most often referred to in text with noun phrases (e.g., computer science) or with proper nouns (e.g., Buffalo Bills). In general, nouns serve as identifiers of text and can be used in natural-language processing systems to identify text samples or portions of samples that are relevant for a particular application or user. Nouns are also used in

[1]Another common variant of the Laplacian matrix can be obtained by replacing L with $I - L$. The eigenvalues change from λ to $1 - \lambda$ but the eigenvectors remain the same.

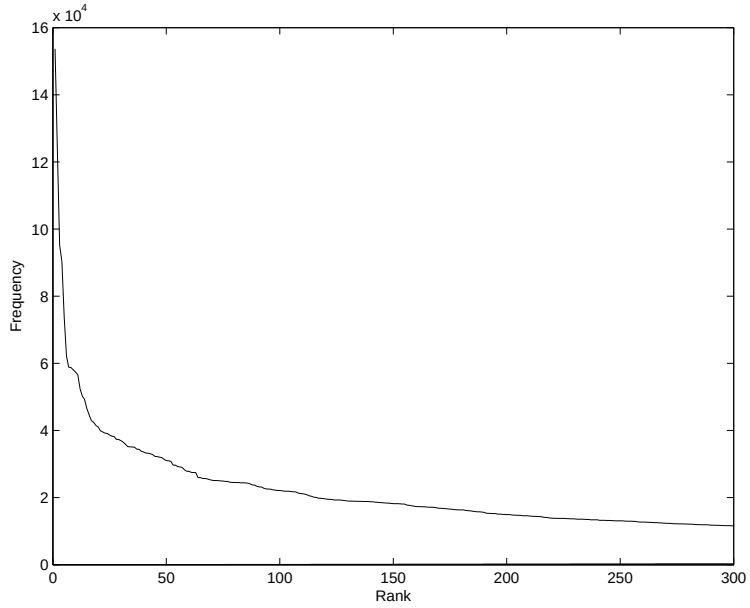

Figure 4: Nouns: Zipf Distribution

advanced information extraction systems for identifying and linking important entities in document summarization and topic detection.

Figure 4 shows the frequency distribution of the top 300 nouns from the BNC sample. Clearly this distribution also obeys Zipf's law.

We use the frequencies from the BNC list to determine whether words occurring in particular pages are more or less frequent than expected, providing a measure of how interesting they may be.

2.2.4 Intuition

Intuitively, any system that seeks to discover novel information must first become aware of what information is already known, and only then look beyond it.

Thus, given an initial query, ATHENS first creates a representation of background knowledge by discovering what is present in the information repository, relevant to the initial keywords. Using this background knowledge, new search queries are created to explore contextually information beyond this first level. Since the goal is to discover information that is richly, though indirectly, connected to the background knowledge, this exploration must be both *highly selective* and *contextualized*.

We now discuss the key steps for novel information discovery and then describe the exact algorithm based on these steps.

Closure: The intuition behind the closure step is to identify the central ideas behind an initial set of keywords and create the background knowledge relevant to it. Closure turns the initial set of keywords into a query, passes them to an underlying search engine, selects the most relevant pages returned, and extracts a concise representation of their contents.

The scope of closure is restricted to information about the initial keywords only. It corresponds to the innermost domain as shown in Figures 1 and 5. The system can now use this background knowledge to contextually probe for information outside this domain.

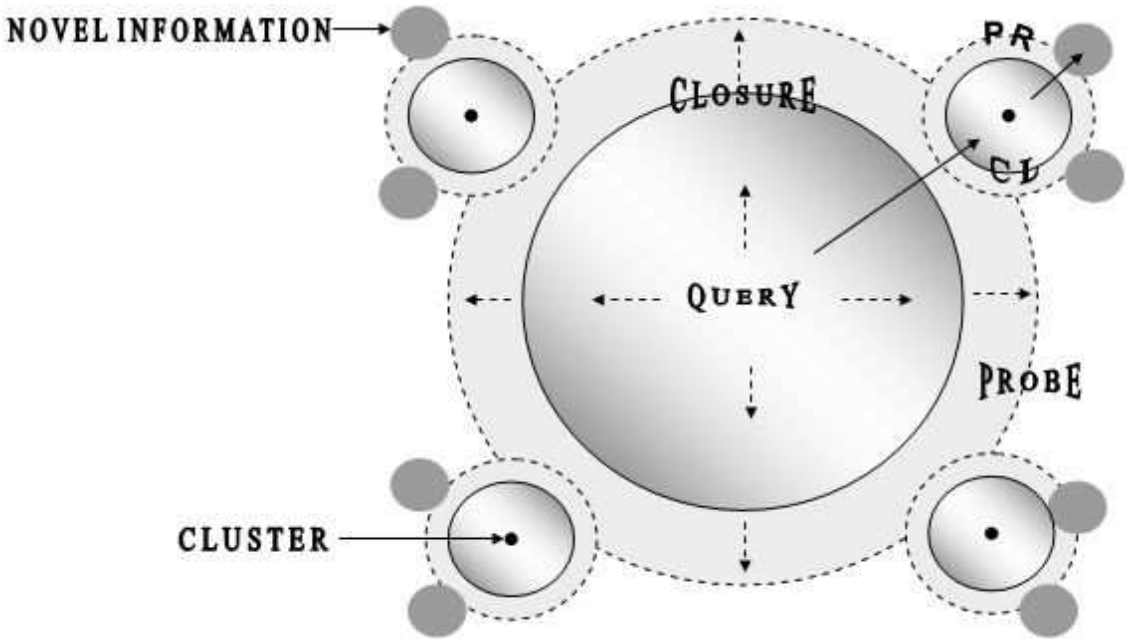

Figure 5: Athens: The Intuition

Probe: The probe step augments closure by creating contextualized new queries, retrieving information from a wider scope. New search queries are generated by combining the original keywords with the important nouns contained in the set of pages retrieved by closure. The original query terms serve to preserve context, while the nouns derived from the closure expand the scope of information retrieved. Since only the most important nouns are used for generating new queries, the search is automatically focused along paths that have greater probability of yielding relevant information.

Cluster: The clustering step has three purposes. First, it reduces an extremely high-dimensional space to a lower-dimensional space in which similar pages are placed in proximity, distance from the origin represents some measure of the interestingness of the relationships of each page to the other pages, and in which the idea of direction is useful in a way that it is not in the original high-dimensional space. Second, it enables the returned pages to be collected into clusters. Third, it enables pages that are not members of clusters to be discarded. The clustering step is the key to searching outwards in a focused and sensible way.

Iterate: The iterate step is a macro step comprising the fundamental steps described above. The closure, probe and cluster steps have retrieved and organized information related to the initial knowledge, but only one step away from the initial knowledge domain.

A set of terms are selected from each cluster, and each set of terms is used as if it were an initial set of terms, and the entire process (closure, probe, cluster) repeated for each one.

We have determined empirically that it is not helpful to repeat the process more than twice, as the contextualization degenerates and the search goes off in unpredictable and useless directions.

2.2.5 Algorithm

Given an initial set of terms Q that is submitted to the system:
1. **Closure**:

(a) Retrieve a subset S of the most relevant web pages for Q using an underlying search engine. Remove any web pages that have already been retrieved. ATHENS uses the Google WebAPI [15] to search over a query and retrieve search results.

(b) Create a list of nouns, along with their frequencies, for every page. This is done by extracting text from each page and tagging nouns using a parts of speech tagger. Remove the nouns present in the original query from each list. ATHENS uses the MontyTagger [10], a parts-of-speech tagger, for tagging the text of every web page.

(c) Combine the noun lists from all pages into a single list of nouns, N. Update the frequency of every noun in N by summing its frequency in each page. This computation gives more weight to nouns that occur most frequently in a set of pages; moreover absolute frequency rather than the relative frequency is considered so that pages containing more content are automatically deemed more important.

(d) Eliminate the less-discriminating nouns as follows: Compute the relative frequency of each noun in the list N. If $f_{n,l}$ and $RF_{n,l}$ represent the frequency and relative frequency of noun n in list l respectively, then the relative frequency of each noun is computed as:

$$RF_{n,N} = \frac{f_{n,N}}{\sum_{i=1}^{size(N)} f_{i,N}} \tag{5}$$

Compute the corresponding relative frequency for each noun in the British National Corpus (BNC) [3]. The relative frequency of a noun in the BNC list is computed as:

$$RF_{n,BNC} = \frac{f_{n,BNC}}{\sum_{i=1}^{size(BNC)} f_{i,BNC}} \tag{6}$$

For nouns that are not present in BNC, the corresponding relative frequency is taken to be zero. Eliminate a noun if the following condition is satisfied:

$$RF_{n,N} < RF_{n,BNC} \tag{7}$$

Equation 7 implies that only nouns with a higher relative frequency in N than in the BNC list are kept. The system also maintains a list of stopwords, containing common English words that are automatically removed from the list of nouns (for example, 'page').

(e) Compute the differential relative frequency, DRF, for a noun n in N as

$$DRF_{n,N} = RF_{n,N} - RF_{n,BNC} \tag{8}$$

Sort N in descending order of $DRF_{n,N}$.

2. **Probe**:

(a) Create new search queries by computing the cartesian product,

$$C = {}^{Q}C_1 * N \tag{9}$$

where ${}^{Q}C_1$ is the set of all possible combinations of terms, that can be formed from Q by removing a single term. Sort C in the descending order of $DRF_{n,N}$.

(b) Select $C_s \ \epsilon \ C$, where C_s comprises the top few queries, representing the most important nouns. Retrieve a fixed subset of most relevant web pages for each query cq where $cq \, \epsilon \, C_s$. Create noun lists per page as in Step 1.

(c) Create a Page–Page Matrix P for all pages retrieved where the (i, j)–th element indicates the relative similarity between pages i and j. If N_i and N_j represent the corresponding noun lists for pages i and j; f_{ni} represents the frequency of a noun n in page i; and $size(S)$ represents the number of elements in set S, similarity between pages i and j can be computed using the Jaccard coefficient [1] as

$$Sim(i,j) \ = \ P(i,j) \ = \ \frac{i \ \bigcap \ j}{i \ \bigcup \ j} \tag{10}$$

where

$$i \bigcap j \ = \sum_{n \epsilon (N_i \bigcup N_j)} min(1, f_{ni}, f_{nj}) \tag{11}$$

$$i \bigcup j \ = \ size(N_i \bigcup N_j) \tag{12}$$

3. **Cluster**:

(a) Compute L, the normalized adjacency matrix of P using the following definition:

$$L(u,v) = \begin{pmatrix} 0 & if \ u \ = \ v \\ \frac{P(u,v)}{\sqrt{d_u d_v}} & if \ u \ and \ v \ are \ adjacent \end{pmatrix} \tag{13}$$

where $d(u)$ is the degree of u, that is the sum of the similarities between page u and all the other pages; and $P(u, v)$ is the similarity between pages u and v. This creates a Laplacian representation of the matrix.

(b) Perform SVD [2, 5, 6] on L to get a k-dimensional vector representation for all pages. Each page is now represented by its corresponding page vector. Compute the dot product between the page vectors.

(c) The following simple clustering algorithm is used to cluster the page vectors: If V_1 and V_2 are two page vectors, they lie in the same cluster if the following conditions are satisfied:

$$\|V_1 + V_2\| \ > \ \alpha \ * \ \|V_1\| \tag{14}$$

$$\|V_1 + V_2\| \ > \ \alpha \ * \ \|V_2\| \tag{15}$$

where α is a constant whose value ranges between 1.8 and 1.92. Insert every new page vector into a cluster, that has the maximum number of its page vectors satisfying the above criteria, when compared with the new vector. For a page vector that is distinct from all existing clusters, a new cluster is created. A geometrical interpretation of page vectors is used to

eliminate those clusters that contain less relevant page vectors (by comparing every page vector with the highly relevant closure page vectors and eliminating if the page vector is geometrically different from all closure page vectors).

(d) For each of the remaining clusters, generate an identifying set of nouns. This creates a set I with one entry for each cluster. ATHENS uses a set of 3 most frequent nouns to identify each intermediate cluster.

4. **Iterate**: Repeat the three previous steps for each cluster, that is, for each $i \epsilon I$ to discover novel information.

2.2.6 Parameters

The algorithm above can be used in a number of different settings, with different parameter values. Depending on the nature of the novel information required, we can restrict the search for novel information to a particular site (*www.ibm.com*), a particular domain (*.uk*) or over the complete Web. The key parameters used in our experiments are listed below. They were chosen empirically, after some experimentation with the system.

- **Number of Pages in Closure**: This parameter specifies the number of relevant pages retrieved during closure, using an underlying search engine. A subset comprising the top 10–20 most relevant pages gives a good approximation to closure in all settings.
- **Number of Pages in Probe**: This parameter specifies the number of relevant pages retrieved for each new query generated during the probe step. As there are multiple queries, the goal is to retrieve only the most relevant information for each new query. Therefore this parameter is usually set to keep the 5 most relevant pages retrieved per query.
- **Number of New Queries in Probe**: This parameter specifies the number of new queries used for probing and retrieving information from a wider scope. The list of nouns generated from closure form a Zipf distribution [16]. This distribution indicates the first 10–25 nouns as the most important. Thus if the original query has two terms, 20–50 new queries can be generated. For restricted settings such as a site search or domain search, usually a value from the lower end of the range, say 20 new queries, is used.
- **Language**: The web pages retrieved using the closure and probe steps are restricted to the English language. As discussed in Section 2.2.1, this system can be adapted to handle information repositories comprising documents in other languages.
- $\boldsymbol{\alpha}$: This clustering parameter essentially represents how much larger the magnitude of the sum of two vectors is, compared to their individual magnitudes. Ideally if the two vectors are exactly the same, the value of α is 2. Therefore vectors are clustered by high values of α. Based on experimentation, a value of α between 1.8 and 1.92 gives a good vector clustering.
- **Cluster Representation**: Experiments performed using ATHENS involve two iterations and a set of clusters is generated after each iteration. The system returns a page of URL's for each cluster. Each intermediate cluster obtained at the end of the first iteration is associated with an intermediate query constructed using the following scheme: A triplet of terms is selected per cluster. Each triplet consists of nouns that occur most frequently with each other, that

is the sum of the pairwise noun–overlap frequency is highest. Triplets give a good enough representation of cluster information without being overly specific and are used as input to next phase.

For representing the novel clusters obtained after the second and final iteration, we use the following scheme: The nouns in each novel cluster are ranked in decreasing order of their frequency in the cluster. The top 15 nouns per novel cluster are selected to represent the contents of that cluster.

The distinction between the above two cluster representations exists for pragmatic reasons. The intermediate queries are only used as input to the next iteration. A word triplet is an optimum size for representing the cluster contents while simultaneously preventing too narrow or too wide a search during the next iteration. However, for representing novel clusters to users, a richer representation of 15 nouns is used.

2.3 System Architecture

This section describes the architecture of the ATHENS system. Figure 6 shows the main components used in the system. We now discuss the role of each component for novel information discovery in the Web:

2.3.1 URL Retriever

The URL retriever accepts a search query and outputs a set of URLs. The URLs correspond to the most relevant pages retrieved for that query using an underlying search engine. The user only needs to specify the number of relevant pages to be retrieved for the input query. ATHENS uses the Google Web API [15], a free web-based service for accessing Google search results from software.

2.3.2 Web Crawler

The web crawler crawls over the retrieved URL's and downloads the corresponding web pages. In other words, this component obtains the information from the web page associated with the URL. ATHENS uses a parallel implementation of this component to crawl and download multiple web pages simultaneously.

2.3.3 Scanner

The scanner component scans the downloaded web pages and extracts the text contained in the page. This component filters out text such as html tags, language code and special characters. The output from this component is the main textual content associated with a particular page.

2.3.4 Tagger

The tagger tags the text outputted by the scanner. Parts of speech tags like those for nouns, verbs and adverbs are used to tag the text. The words tagged as nouns are selected while the other words are filtered out. The reason for this is that nouns are the best identifiers of information and can be used to identify the content associated with a web page. The output of this component contains the set of nouns along with their frequency in each page. ATHENS uses MontyTagger, a parts-of-speech based tagger for tagging the text.

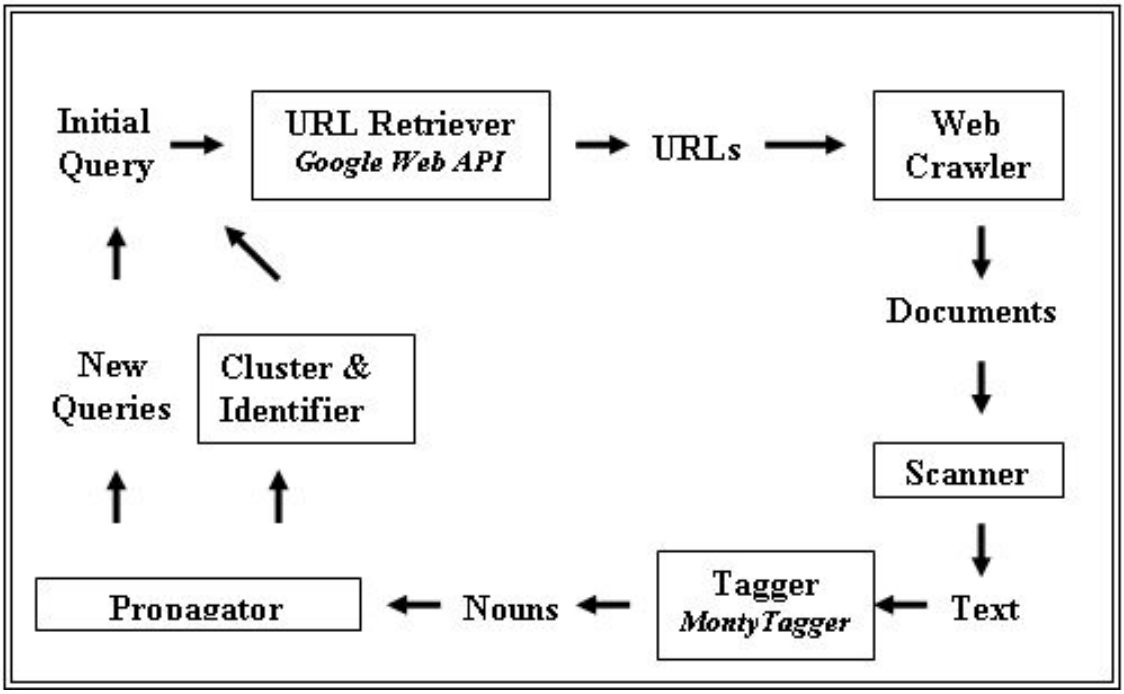

Figure 6: ATHENS System Components

2.3.5 Propagator

This component is used at the end of the closure step to combine nouns from all retrieved pages into a unique list. The nouns are sorted in the descending order of their net frequency. A list of stop words is maintained to filter out the very common English words. The nouns are then reranked using their relative frequency in the BNC list (see Section 2.2.5). A list of new queries, to be used for probing, is generated by combining the query with the top few nouns.

2.3.6 Cluster & Identifier

This component is used at the end of the probe step to organize information efficiently and eliminate the less relevant information. The pages are organized into a page–page matrix containing matrix similarities (see Section 2.2.5). This matrix is decomposed through the process of Singular Value Decomposition into the product of matrices containing singular values and singular vectors. This technique is then used to cluster the pages and eliminate less relevant pages as discussed in Section 2.2.5. Finally, a set of nouns is generated per cluster to identify the contents of the cluster.

2.4 Tools and Software

2.4.1 J2SE 1.4.2

J2SE or Java 2 platform standard edition [9] is the complete open source Java programming environment from Sun Microsystems inc. The advantage of programming in this environment is that it is platform independent. The complete ATHENS system is coded in Java to avoid interdependency on any other tool.

2.4.2 Google Web API

The Google Web API [15] is a free web-based service provide by Google. It allows the search results of Google to be accessed directly from a programming environment. This service gives query access to Google's Web search, enabling the development of software for accessing billions of Web documents that are constantly refreshed. Developers can issue search requests to Google's index of more than 4 billion web pages and receive results as structured data. Google Web APIs support a richer search syntax than the *www.google.com* site.

2.4.3 MontyTagger v1.2

MontyTagger [10] is a rule–based, part of speech tagger. It annotates English text with parts of speech information, for example "dog" as a noun, or "dog" as a verb. The tags used by MontyTagger are from the Penn Treebank tagset [12]. Commonsense knowledge is used for tagging the words in text. The system also uses a trained lexicon for the tags in Penn Treebank and rule files. A tokenizer for English is included to tokenize text and other tools are used to evaluate performance.

2.4.4 JAMA: Java Matrix Package

JAMA [11] is a basic linear-algebra package for Java. It provides user-level classes for constructing and manipulating real, dense matrices. JAMA is comprised of six Java classes: Matrix, CholeskyDecomposition, LUDecomposition, QRDecomposition, SingularValueDecomposition and EigenvalueDecomposition. These matrix classes provide the corresponding fundamental operations of numerical linear algebra.

3 Experiments

We have performed a number of experiments across a number of domains to validate the utility of our system for discovering novel information. To enable a consistent evaluation of results, the parameters across each experiment have been kept the same.

3.1 Setup

We arrived empirically at an effective set of parameter values that best lead to the discovery of novel information. The following is the list of parameter values used for conducting each experiment:

Iteration 1

- **Number of Pages in Closure**: Closure is restricted to the top 10 most relevant pages retrieved for the query using an underlying search engine.
- **Number of Pages in Probe**: Probing is restricted to top 5 most relevant pages retrieved for each subsequent query generated.
- **Number of New Queries in Probe**: The top 25 new queries are used for the probing stage.
- **Language**: English

- $\boldsymbol{\alpha}$: is set to 1.92 for the first iteration.
- **Intermediate Queries**: For representing each intermediate cluster, an intermediate query comprising a triplet of nouns is used. Each intermediate query serves as input into the next iteration of the experiment. If C_I is the representation for an intermediate cluster, then

$$C_I = (noun_1\ noun_2\ noun_3) \tag{16}$$

  where the noun triplet consists of nouns with each pair occurring together most frequently, that is sum of the pairwise overlap for all pairs is maximum.

Iteration 2

- **Number of Pages in Closure**: Closure is restricted to the top 5 most relevant pages retrieved for each cluster query using an underlying search engine.
- **Number of Pages in Probe**: Probing is restricted to top 5 most relevant pages retrieved for each subsequent query generated.
- **Number of New Queries in Probe**: The top 25 new queries are used for the probing stage.
- **Language**: English
- $\boldsymbol{\alpha}$: is set to 1.87 for the second iteration.
- **Novel Cluster Labels**: For representing each novel cluster, we use a set of 15 nouns. If C_N is the representation for a novel cluster, then

$$C_I = (term_1, ..., term_{15}) \tag{17}$$

  where the terms represent the top 15 nouns with the highest frequency in the cluster.

3.2 Quality Measures for Results

The results of each experiment are a set of novel information clusters. Each cluster consists of a listing of the clustered web pages and is represented in the output by a 15-word set of terms that best describe the content of that cluster. The label terms can be used as a search query to retrieve the web pages corresponding to the cluster. The system creates a web page with links to the URLs of each cluster as part of its output, but these are too large to present.

Clusters are rated using two criteria: coherence (how well does the cluster reflect related content), and novelty (how well does the cluster reflect our goals for novel information).

3.3 Problems of Evaluation

There are three fundamental problems associated with any technique for evaluating novel information. These problems are discussed in the following subsections.

3.3.1 Not Knowing What We Don't Know

Since the system finds information that is not directly connected to the initial terms as starting point, it is hard to determine if relevant or appropriate information has been missed. This has been addressed in two ways. First, we used queries suggested by subject experts who were knowledgeable enough to guess what kind of indirect information ought to be discovered (although this is somewhat problematic because even subject experts are not used to thinking about indirectly connected information). Second, we carried out an experiment (reported elsewhere [13]) in which we ran ATHENS as it would have run on September 12th 2001. Since the background to al Qaeda and the World Trade Center attacks has been extensively studied, we were able to show that almost all of the discovered information was successfully found by ATHENS.

3.3.2 Ontological Bias

This bias refers to the fact that humans expect novel information relative to a topic to fall into a similar part of their mental ontology. In other words, given a query in area X, the expectation is that the corresponding novel information should also lie in area X. For instance, if the initial query is *Multiple Sclerosis*, a user typically expects novel information about topics such as *neurological disorders* or *mental illnesses* which fall under the same ontological relationship (that is, health). However, there also exists novel information about other areas, for instance *rehabilitation* or *charitable institutions* that is strongly related but does not fall in the expected ontology. Humans may therefore categorize this information as irrelevant when in fact it is strongly connected to the initial keywords.

3.3.3 Cluster Description Problem

This problem arises as a consequence of the cluster labelling process. Each cluster is labelled using the most frequent terms associated with the cluster. This labelling scheme describes the content of the cluster very well. This is illustrated by the fact that a search on these terms retrieves useful novel information about a specific topic. It is also consistent with current search technology which uses individual terms rather than an ontology-based search.

Humans tend to form and recognize an ontology-based representation of information more easily. While the labelling scheme is an accurate representation of the cluster contents, it may not always produce a cohesive representation of a topic. However, the nature of the terms in the cluster labels is not based on ontology but on how information about the cluster contents can be accessed in the best possible manner. A search using the terms in a cluster label will retrieve a set of web pages that best describe the content of the corresponding cluster.

Further, the Web itself is not organized as an ontology. The existing search technology does not rely on any notion of semantics as a mechanism for retrieving information. This suggests that an ontology-based search may not be necessarily good for retrieving the contents of a cluster.

3.4 Evaluation Techniques

There exist useful techniques, such as Maximal Marginal Relevance [4], that can be used for ranking documents by their novelty value relative to a query. However, existing techniques can only be used to measure the novelty of documents retrieved directly in response to a query. They cannot be

used for evaluating novel information, as this information is indirectly related to the original query topic.

It is also clear that users' perception of novelty depends on their entire background knowledge, not all of which may be captured by an initial set of keywords. Hence, some of the information returned by ATHENS may not be novel to the user. The threshold of satisfaction experienced by different users is likely to be different.

We have devised a set of techniques for validating the results of experiments. These techniques are described in the following subsections.

3.4.1 Output Comparison

For certain experiments, it is useful to compare results of similar initial queries in order to validate them qualitatively. The same experiment can be conducted across separate domains (for instance, a *site search* across two web sites or a *domain search*) and the results compared. Such a comparison itself acts as a validation measure because the differences in results should be explainable in terms of the differences in environment. Such experiments are also useful to compare similar concepts in different settings. Therefore we have conducted the first experiment as a site search across separate domains and evaluated the results by comparison.

3.4.2 Expert Judgement

For the remainder of the experiments, we have employed a subject expert to assess the quality of novel information retrieved in the corresponding knowledge domain. The expert was asked to draft an *expectancy list* of the topics that were expected to be novel with respect to a particular initial query. The expert compared the result clusters (indicated by the labels) with the expectancy list to evaluate the results. Each novel cluster was then assigned to one of the following 5 categories:

- **Relevant (R)**: refers to information on the expectancy list of the expert, in other words, expected novel information.
- **Partly Relevant (PR)**: refers to information that is partly relevant or novel.
- **Irrelevant (I)**: refers to irrelevant information, that is information which cannot be classified as novel.
- **Not–Thought–Of (!)**: refers to information that is novel, but was not on the expert's expectancy list and should have been.
- **Not–Known(?)**: The expert is unable to judge the information for novelty based on the labelling information available.

4 Results

The results of each experiment are a set of novel information clusters. Each cluster is represented by a label comprising a set of terms that best describe the content of that cluster. The label terms can be used in a search query to retrieve the web pages corresponding to the cluster.

The results are presented both as a *table* and as a *graph*. The two modes of result presentation are described as follows:

4.0.3 Result Table

The result table contains a set of intermediate queries, each associated with a set of labels that represent the novel information obtained. The exact role played by these queries and labels is discussed below:

Intermediate Query: These queries represent the intermediate clusters that are obtained after performing the first iteration of each experiment. Each of these queries is a set of 3 nouns that are used as input queries for performing the second iteration of the experiment. Including these queries in the final result is helpful because it identifies the path in which the search for novel information proceeds.

Novel Cluster Labels: These labels represent the novel clusters that are obtained from the second level iteration in an experiment. Each of these labels comprises a set of 15 nouns that identify the content of the corresponding novel cluster.

The main purpose of the labels is to provide a representation of the contents of the novel clusters. These labels are used to judge the quality of novel information using the evaluation criteria described in Section 3.4.

4.0.4 Result Graph

The result graph depicts the novel clusters in 2–dimensional space. Each cluster is represented by a cluster vector which is the mean of the page vectors for all pages in that cluster. The cluster vectors are truncated and plotted in 2 dimensions to visualize the inherent underlying relationship between the clusters. Further, as there exist a set of novel clusters for each intermediate query, this graph also displays the amount of coherence among each set of novel clusters.

4.1 Queries

We have performed 5 experiments, with the first experiment being a combination of two experiments. Each experiment explores a different knowledge domain in the Web. Some of the experiments have been conducted in restricted search domains, for instance restricting the novel information search to a particular site (*www.ibm.com*) or a particular domain (*.uk*).The following is the list of queries used for performing the experiments:

Data Mining
This query was used to perform a site search experiment. The search was restricted to the corporate web sites of two organizations: IBM and Microsoft, that is *www.ibm.com* and *www.microsoft.com* respectively. The motivation for this experiment is to explore the organizational relevance of each research field within IBM and Microsoft. These results could be understood as measuring how well data mining has been integrated into each organization; or, alternatively, how well each organization is positioned to exploit significant progress in data mining.

"al qaeda" "bin laden"
This experiment was conducted across the complete Web. The motivation behind this experiment is to explore the huge reservoir of information related to terrorism and counterterrorism and discover novel information. This experiment illustrates the use of phrases as search terms.

Case Grammar
This query was used to conduct the search for novel information across the complete Web. The motivation behind this query is to explore the linguistic knowledge domain and ascertain the quality of the novel information obtained.

Multiple Sclerosis
This experiment was conducted across the complete Web. A search on this query typically retrieves information associated directly with the disease. The goal of this experiment was to explore information about the health, social, research and humanistic factors associated with this disease that is not obvious and hence not directly retrievable.

Multiple Sclerosis (.uk)
This experiments repeats the previous experiment but with the search restricted to a particular domain, in this case *.uk*. The motivation behind this experiment is to obtain a domain wide information about this topic and compare the results with those of the previous experiment. This experiment also gives interesting insights into the organization of information in the Web, that is how it is organized in a specific domain as compared to the whole Web.

4.2 Results for the query: data mining

The goal for this experiment is to discover novel information related to data mining within the organizations IBM and Microsoft. The search is restricted to the corporate web sites: *www.ibm.com* and *www.microsoft.com*. The results in tabular and graphical form are presented below. The tabular form comprises each intermediate query (**IQ**) along with the 15–noun labels for the associated novel clusters.

Table 1: IBM Cluster Labels

IQ1	**Tivoli, Storage, Manager**
1	FAStT, Windows, Host, Adapter, WebSEAL, Linux, Server, SPNEGO, Management, Software, Plug, Guide, Access, User, Support
2	Management, Software, Support, Business, Solutions, Retention, Optimization, Resource, Products, Download, Services, Edition, Technical, Extended, Protection
3	Software, Support, Resource, Flashes, APAR, Information, Databases, FAQs, Manuals, Submit, Redbooks, Strategy, Alerts, Solve, Technotes
4	NetView, OSIBM, Software, Management, Business, Support, Access, Information, Performance, Resource, System, ExchangeIBM, Decision, DatabasesIBM, Network
5	WSBCC, Policy, Access, Program, Files, WebSphere, Director, WebSEAL, J2EE, Migration, Teller, Authorization, Server, Java, Temp
IQ2	**OLAP, Server, Services**
6	Windows, Database, Integration, Application, Guide, Manager, ODBC, FixPak, Websphere, Information, Essbase, Miner, Analysis, Installation, Hybrid

Table 1: IBM Cluster Labels (continued)

7	TotalStorage, WebSphere, Essbase, Software, Grid, Support, Java, Guide, Release, Volume, Controller, Windows, Editor, Open, Table
8	Windows, Universal, Database, Extender, Software, Cube, Integrated, Administration, Programming, Support, Guide, Edition, Information, Solutions, xSeries
9	WebSphere, Business, Support, Management, Enterprise, Technical, Information, Application, Studio, Software, Products, Partners, Performance, Industries, Training
10	Java, Technology, Eclipse, Application, Toolkit, Tool, Computing, Development, MPEG, Technologies, Interface, Management, Framework, JDBC, Grid
11	HTML, Windows, Managing, GC88, EnglishPDF, Management, SC13, Facility, SC88, PortuguesePDF, S517, SC12, SA30, SC10, SC27
IQ3	**OLAP, Cube, Edition**
12	Hyperion, Server, Installation, iSeries, Corporation, Guide, Copyright, Rights, Portions, Reserved, November, Document, Solutions, Number, GC18
13	WebSphere, Rational, Lotus, Tivoli, Studio, Server, developerWorks, Java, Business, Windows, Toolbox, Support, Vice, President, Information
14	Business, Management, Software, Windows, Support, United, States, Products, Facility, Query, Database, Advantage, Multiplatforms, Warranty, WebSphere
15	FixPak, Software, Support, Windows, Business, Database, Group, Neutral, Developers, Products, Partners, Technote, Problem, Solution, Universal
IQ4	**Business, Intelligence, Services**
16	WebSphere, Linux, Server, Lotus, Application, Software, Support, Tivoli, Management, Products, Integration, Enterprise, Information, Training, Certification
17	Alphablox, Management, Software, Corporation, Support, Information, Submit, Warehouse, Presse, FAQs, Strengthens, Espace, Acquire, Contacts, Actualits
18	Training, Linux, Global, United, States, Application, Certification, Development, Support, Education, Industries, Products, Learning, Conferences, Partners
19	Education, Norge, Hjem, Produkter, Tjenester, Support, Download, Velg, Utdanning, Kurskatalog, Kurssk, Sertifiseringer, Spesielle, tilbud, informasjon
20	Tivoli, Certification, Certified, Professional, Exam, FAQs, Training, Partner, Redbooks, Certifications, Preparing, Support, doesnt, Education, Sylvan
21	Information, Management, Informix, Training, Global, Education, Database, United, States, Certification, Conferences, Technical, General, Bird, Discover
IQ5	**OLAP, Server, Release**
22	Informix, Software, Products, Support, Dynamic, Business, Information, Product, Systems, Advantage, Warranty, Management, States, United, Solutions
23	Sametime, WebSphere, Drive, Software, Domino, Boscov, Support, Services, Business, Lotus, Advanced, Tools, Informix, Application, Adapter
24	Application, WebSphere, Java, Software, Upgrade, Support, Services, Servers, Independent, Policy, Group, Neutral, Product, Edition, Recommended
25	Tivoli, Directory, Software, WebSphere, Support, Product, Flashes, APAR, FAQs, Manuals, Submit, Strategy, Redbooks, Portal, Patch

Table 1: IBM Cluster Labels (continued)

IQ6	**OLAP, Server, Integration**
26	WebSphere, Business, Host, Siebel, Systems, Access, Support, Solution, Client, Software, Services, Publisher, Management, Products, Application
27	Domino, LANSA, Software, Websphere, Systems, Enterprise, Solutions, Corporation, Application, Enhancements, Jacada, Java, Design, Business, Lotus
28	WebSphere, Business, Lotus, Domino, Grid, Foundation, Software, Support, MicroStrategy, Products, Windows, Application, Services, Document, Toolkit
IQ7	**Intelligent, Miner, Software**
29	Tivoli, Orchestrator, Business, WebSphere, Integration, Orchestration, Support, Products, Windows, Provisioning, Warranty, Product, Scoring, Solutions, Manager
30	Support, Information, Management, Business, Advantage, Products, Product, Visualization, Solutions, Modeling, Download, Scoring, United, States, eServers
31	Tivoli, Lotus, Server, WebSphere, Manager, Storage, Everyplace, Selecione, business, Management, Informix, Application, Family, Voice, Solutions
IQ8	**Clinical, Genome, Miner**
32	TGen, Services, Linux, Business, Global, Kyoto, National, Cluster, Hospital, Kitaoka, Institute, Research, Storage, Database, Support
33	Research, EMBL, European, University, Centre, Bioinformatics, Institute, Molecular, Life, Sciences, Biology, Hutchison, Cambridge, LABORATORY, Healthcare

Table 2: MSN Cluster Labels

IQ1	**Server, Services, Analysis**
1	Windows, Download, Security, SP3a, Excel, Database, Update, Edition, Updates, Terminal, Systems, Office, Slammer, Exchange, ApplicationsSystem
2	Windows, Site, Information, Management, Edition, Report, Excel, Update, Office, Features, Enterprise, Security, Standard, Terminal, Software
3	OLAP, Components, Report, MSDE, Download, Database, Windows, Downloads, Site, Information, Manager, Support, Technologies, Technical, Data
4	MOLAP, Windows, Wizard, ROLAP, Aggregation, Usage, HOLAP, Design, Storage, Optimization, Partition, Manager, Query, Cube, Terminal
5	Windows, Security, Monitor, Performance, Update, Query, Time, ROWS, Data, Network, Manager, Download, System, Oracle, Internet
6	Monitor, Network, Windows, Domain, EnvironmentSee, Performance, Application, Subsystem, Implications, Processor, RAID, Maximize, Throughput, Manager, Bytes
7	Windows, Client, Manager, Tool, Alpha, Trace, Syntax, Tools, Systems, Processors, CLEANER, Data, Control, Site, VIEWER
IQ2	**OLAP, OLTP, Server**
8	Data, Warehouse, Analysis, Services, Information, Business, Design, Corporation, Mining, Architecture, Site, Support, Hyperion, Essbase, Microsoft

Table 2: MSN Cluster Labels (continued)

9	Planning, Windows, Design, Technical, Resources, Licensing, Designing, Data, Warehouse, Database, Site, Application, Partners, Technologies, Community
10	Windows, Datacenter, Edition, Limited, Services, OEMs, January, Enterprise, October, Free, March, Magazine, Intelligence, Information, Partners
11	Accelerator, Services, Technologies, Analysis, Data, Resources, Windows, Component, Site, mart, Technical, Transformation, Worldwide, Support, Partners
IQ3	**Storage, Server, Access**
12	Windows, SharePoint, Business, Framework, Portal, Services, Site, Office, Network, Team, Corporation, Internet, Series, Redmond, Enterprise
13	Windows, Defragmenter, NTFS, Exchange, System, Active, Directory, Paging, Volume, Cluster, Paul, Contact, Corporation, Trademarks, Statement
14	NTFS, Windows, System, Technical, Reference, FAT32, Operating, Disks, Volumes, Dynamic, Technologies, Service, Compatibility, Copy, Dependencies
15	Yukon, BizTalk, MathTutor, MessageBox, Studio, CREATE, AddNumbers, Belux, Framework, User, Schema, ProductID, ASSEMBLY, Data, Messaging
IQ4	**Cube, CREATE, Store**
16	Project, Server, Windows, Business, Services, Professional, SharePoint, Site, Standard, Solutions, Team, Excel, Information, Database, Report
17	Project, Server, Exchange, Windows, Professional, Smart, System, Network, Cluster, Services, Internet, Messaging, Instant, Service, Site
18	Directory, Active, Server, Windows, Cingular, Data, OLAP, Describe, Media, Query, English, OutlookSoft, Services, Excel, Module
19	Project, Windows, Server, SharePoint, Services, Internet, Central, Analysis, Products, Technologies, Sigma, Teradata, Portal, Database, Professional
20	Server, Data, Analysis, Intelligence, Office, Business, Smart, Portal, OLAP, Instructor, Completion, Certified, Professional, Exams, Prerequisites
IQ5	**CATIA, Server, Code**
21	SharePoint, Portal, Support, Windows, Development, Software, Services, Reference, Advanced, Guide, Products, Versions, Downloads, Download
22	Site, Windows, Corporation, Commerce, UNIX, Edition, Rule, Builder, Management, TechNet, Active, Directory, Migration, ACLs, DTCs
23	Windows, UNIX, Interix, Perl, Application, Services, Migration, Corporation, Migrating, Custom, Control, System, ActivePerl, Win32, Rewriting
24	Windows, UNIX, SIFAgent, Site, Resource, Office, Critical, Edition, Messenger, Deployment, Commerce, Error, Branding, Solution, Toolkit
25	Exchange, Windows, Directory, Membership, MAPI, UNIX, Site, User, Internet, Forms, LDAP, Research, Services, Active, Corporation
26	Windows, Directory, LDAP, Office, Membership, Setup, Active, Connector, MSADC, ACLs, Installer, CedarBank, Agreement, Site, COMTI
27	Application, Block, Management, Portal, Framework, Reference, Architecture, SharePoint, Caching, NETA, Exception, Enterprise, Data, Windows, User

Table 2: MSN Cluster Labels (continued)

28	DHCP, Active, Directory, Windows, Description, Address, Authorization, Host, Time, Analyzing, Audit, domainThe, Event, Restarting, MYDOMAIN
IQ6	**SharePoint, Server, Windows**
29	Deployment, Services, Download, Resource, Internet, Free, Worldwide, Technical, Edition, Deploying, Security, Information, Reference, Kits, 2003Windows
30	Services, Download, Administrator, Managing, Security, Site, Part, Configuration, Service, Edition, MediaDriversOffice, ApplicationsMobile, DevicesMacintosh, PlatformsServer, ApplicationsSystem
31	Portal, Resource, Services, Internet, Explorer, Office, Team, Edition, Product, Kits, Deployment, Document, Professional, Information, Products
32	Services, Portal, Products, Application, Technologies, Site, Security, Internet, Information, Office, Manager, Tool, Resources, Directory, FrontPage
33	Services, Products, Portal, Business, Technologies, Office, Migration, Corporation, Security, Systems, Worldwide, Management, Contact, Free, Statement
34	Portal, Security, Download, Explorer, Internet, Update, Edition, Services, Office, Configuration, Directory, Analyzer, Client, Active, Site
35	Services, Download, Deploying, Deployment, Office, Site, Migration, Security, Information, MediaDriversOffice, ApplicationsMobile, DevicesMacintosh, PlatformsServer, ApplicationsSystem, Management
36	Security, MCMS, Products, Portal, Download, Technologies, Services, Management, Office, Connector, Update, Migration, Document, Library, ToolsDevelopment
37	Management, Services, Security, Update, Product, Download, Technology, Community, UNIX, Technical, Service, Downloads, TechNet, Patch, Products

Table 1 identifies the IBM novel clusters with the labels for each cluster. Clearly, the IBM results show tremendous diversity in the kind of information retrieved, such as information about *Websphere* for business management; *bioinformatics*; *OLAP Cubes* for business intelligence and financials; *Domino Lotus Software* for Web usage Mining; *Websphere Express Portal* for product information management; *Tivoli platform* for business oriented storage services, among others.

Figure 7 is the two–dimensional plot obtained after SVD on the novel clusters obtained from IBM, truncating to two dimensions. The groups of novel clusters show good coherence, for example the novel clusters pertaining to *OLAP Server Services*.

Table 2 contains the corresponding cluster information for Microsoft. There is an interesting overlap of research areas common to both organizations, for example OLAP cubes for business intelligence. The data for Microsoft reveals the existence of several interesting solutions like *OLAP and OLTP models* for data mining and warehousing; *Accelerator for Six Sigma* for financial analysis; *Sharepoint Portal Server* and *SQL Data Analysis Server* for business intelligence, among others. However, the novel information is not as diverse as that for IBM.

Figure 8 is the two-dimensional plot of the subclusters obtained from Microsoft. The plot exhibits overlapping clusters which indicate an underlying relationship for the novel information retrieved. Most of the solutions provided are packaged into a unified platform such as the SQL Server platform or Project Server platform.

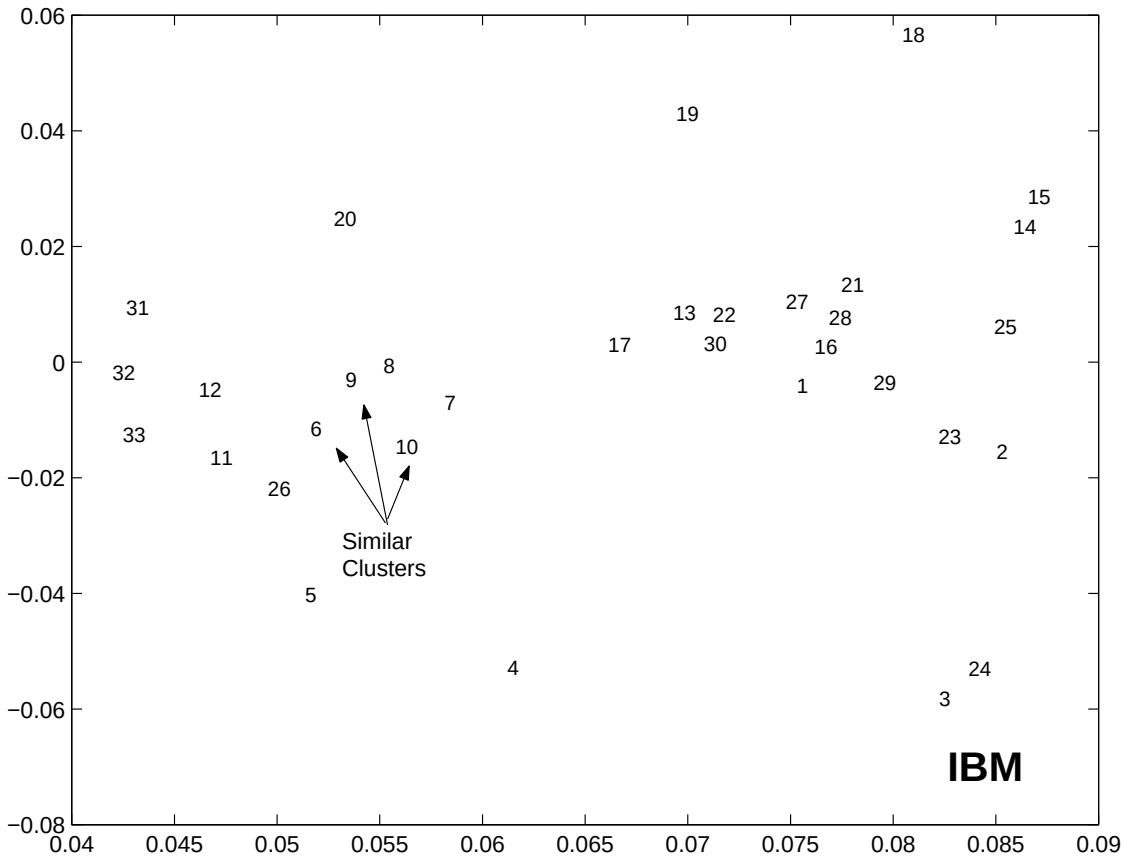

Figure 7: Novel Information Clusters

Clearly, the results for IBM exhibit a high degree of quality and coherence in the nature of information retrieved. They imply a strong integration of the data mining work done in IBM with other key areas of research and development, emphasizing good coordination. For instance the results show that IBM provides efficient hardware backup (*FAStT Storage Server*) to critical business, bioinformatics and data mining applications. The results for Microsoft, however, are less diverse but overlap extensively. This implies that while some data mining research has been successfully incorporated in key software platforms, the organizational philosophy is not fully directed towards this field. Further details of this experiment can be found in [14].

4.3 Results for the Query: "al qaeda" "bin laden"

The goal for this experiment is to discover novel information relative to the query *"Al Qaeda" "bin Laden"*. The search for novel information is performed within the complete World Wide Web. The results have been rated by a subject expert according to the categories described in Section 3.4.2.

Table 3: "al qaeda" "bin laden" Cluster Labels

IQ1	**Iraq, Bush, Saddam**	
1	United, Clarke, States, Security, Council, Baath, Hussein, President, Arab, White, House, Nations, American, East, Middle	R

Table 3: "al qaeda" "bin laden" Cluster Labels (continued)

2	United, Security, Nations, Council, President, Hussein, America, States, September, February, January, December, March, Middle, November	R
3	Hussein, President, MSNBC, House, ABCNEWS, Rumsfeld, White, Kerry, Prime, Minister, International, Business, ENTERTAINMENT, TECH, Newsletters	PR
4	House, White, State, Union, July, Tenet, American, WSWS, Africa, Niger, Hussein, Rice, Security, United, States	R
5	White, House, Suskind, Kerry, Paul, American, Cabinet, Treasury, International, SERVICES, Education, Tools, Alerts, SEARCH, George	R
6	Kerry, America, President, Security, Libya, American, PLAY, Report, International, Hussein, Nader, United, SPECIAL, SERVICES, World	R
7	Security, Council, Presidential, Statement, UNSCR, American, United, IAEA, March, America, November, President, Nations, October, Hussein	R
8	Middle, East, President, America, Reagan, November, Endowment, Freedom, Ronald, National, Europe, Democracy, United, States, January	PR
IQ2	**Saudi, Afghanistan, Islamic**	
9	Arabia, International, University, Secretary, Pakistan, Organization, Information, Embassy, King, United, Council, Assistant, Craner, Office, Washington	PR
10	Taliban, Pakistan, INDIA, United, Kabul, Muslim, BUSH, Iraq, Arabia, Muslims, States, Arab, Iran, HOUSE, Unger	R
11	Property, Arabia, Riyadh, Office, Dubai, Middle, Arab, Dana, International, Firm, Fraih, Jordan, Yemen	R
12	Dostum, Rashid, Taliban, Muslims, Karzai, Medieval, Abdul, Thursday, Muslim, Hamid, JNMP, Prophet, Sharif, Shah, Mazar	R
13	Arabia, International, Women, Amnesty, Convention, Commission, Africa, Kingdom, Torture, March, January, Human, Rights, Minister, Najran	R
14	Kashmir, Sudan, Muslim, Terrorism, Harakat, Taliban, Sunni, Yemen, Muslims, United, States, Arabia, Americans, Printer, Friendly	R
IQ3	**Islamic, Qaida, Jihad**	
15	Bosnia, Serbs, Kosovo, Macedonia, Albanian, Afghanistan, Bosnian, Balkans, Muslim, Destro, International, Muslims, BlackJade, GREECE	R
16	Macedonia, Zawahiri, Kosovo, Ayman, IIRO, Macedonian, Relief, International, Mohammed, Zawahri, Saudi, Osama, Albanian, Organization, Dnevnik	R
IQ4	**Afghanistan, United, States**	
17	Taliban, Iraq, Department, Armitage, Kabul, USAID, Madagascar, Development, Mission, Health, Contact, Information, National, Report	R
18	Archive, UNDP, July, Population, Trade, Excel, General, National, Information, Iraq, Department, Project, Resident, Consulate	I
19	Taliban, Information, ARMY, July, Marine, National, Navy, Service, Corps, Pacific, Rights, President, WASHINGTON, Marines, Force	PR

Table 3: "al qaeda" "bin laden" Cluster Labels (continued)

20	Kabul, Pentagon, Karzai, Women, Taliban, Herold, December, October, Civilian, Dossier, Marc, American, Bush, Aerial, Bombing	R
IQ5	**Iraq, Hussein, Rumsfeld**	
21	Saddam, United, Bush, States, Iran, Donald, President, Defense, Secretary, Washington, Baghdad, Gulf, Reagan, American, Kuwait	R
22	Bush, Tenet, President, Saddam, Zarqawi, United, Marines, Saddams, Fallujah, Marine, Tuesday, States, America, MSNBC, January	R
23	Monday, MSNBC, Ghraib, Defense, Newsweek, Bush, Qaida, Department, Donald, Secretary, Yorker, Powell, House, State, Abuse	R
IQ6	**Saudi, Osama, Arabia**	
24	MSNBC, Qaida, Press, Business, Iran, TERRORISM, Johnson, Bush, Security, Interior, Ministry, Newsletters, Section, Pakistan, Nightly	PR
IQ7	**Bush, Carlyle, Saudi**	
25	Arabia, House, George, American, Group, Baker, United, President, Washington, White, Osama, Iraq, September, James, Prince	PR
26	Group, Frank, Carlucci, BuzzFlash, Baker, Secretary, Defense, George, James, September, Washington, Rumsfeld, President, Perspectives, Archive	PR
IQ8	**Saudi, Qaida, Friday**	
27	Johnson, Aljazeera, Riyadh, Iraq, Arabia, Contact, American, Arabian, Arab, Muqrin, Site, Guide, Culture, Tools, Email	R
28	Johnson, Arabia, Muqrin, Riyadh, Aljazeera, Saturday, Jubeir, American, International, Paul, Embassy, Edition, SEARCH, Iraq, Westerners	PR
29	Johnson, Arabia, Aoofi, Hawaly, Thursday, Riyadh, Press, ABCNEWS, Paul, TECH, World, July, Internet, Copyright, Chief	PR
30	Johnson, Arabia, Post, Press, Hawaly, Aoofi, Paul, Thursday, Riyadh, True, Capitalist, Chief, American, Amnesty, Moqrin	PR
IQ9	**Pakistan, Bush, Taliban**	
31	Afghanistan, Islam, Islamic, Terrorism, Kashmir, Afganistan, Rights, United, Asia, Saudi, American, India, September, York, Central	R
32	Enron, Afghanistan, UNOCAL, Cheney, India, President, Ahmed, General, Clinton, Washington, Halliburton, United, Indian, CentGas, Karzai	R
IQ10	**Terrorism, Osama, August**	
33	Ladin, Afghanistan, Sudan, Islamic, American, Sudanese, Ladins, Arab, Saudi, Khartoum, United, States, Veterans, Bombing, Soviet	R
34	Pakistan, Afghanistan, Sudan, Arab, Islamic, Taliban, United, American, Saudi, States, Clinton, September, Israel, Terrorist, Americans	R
35	Islamic, Jihad, Shqaqi, Gaza, Strip, Palestinian, Israel, Territories, Egypt, Jerusalem, Muslim, Aqsa, Tamimi, October, Aviv	R
36	Islamic, Mujaheddin, Saudi, Pakistan, Afghanistan, Idaho, Peshawar, Taliban, Observer, Casey, Stinger, Soviets, Americans, Khost, Arab	R
IQ11	**Military, Organization, Espionage**	

Table 4: Expert Validation Statistics

	Number of Clusters	%age
Relevant	29	63
Partly Relevant	12	26
Irrelevant	4	9
Unknown	1	2
Total	46	100

Table 3: "al qaeda" "bin laden" Cluster Labels (continued)

37	INTELLIGENCE, Agency, Security, National, Commission, Central, COMMUNITY, Department, Defense, president, Operations, Director, Congress, Council, Government	R
38	Intelligence, Defense, Security, National, United, American, Force, States, Agency, Government, Army, Department, Office, Cold, Operations	R
39	NATO, World, United, Fact, Germany, States, France, Union, Security, Encyclopedia, USSR, Francis, Soviet, Spain, Gary	R
40	Toyama, Mitsuru, Japan, Black, Dragon, Japanese, Korea, Motojiro, Akashi, Kempei, Spymaster, Russian, Kyoshisha, Army, Russo	?
IQ12	**Harthi, East, Middle**	
41	Saudi, Arabia, Islamic, United, States, Afghanistan, Saudis, American, Muslims, Ahmed, Muslim, Fawzan, America, Washington, Islam	PR
42	Saudi, Minister, Prince, Abdul, Aziz, Israel, President, Arabia, Prime, General, Deputy, Kingdom, King, Fahd, Iraq	PR
43	Saudi, Arabia, Iraq, Arab, United, Water, States, World, Saudis, Jeddah, AMERICAN, Israel, Environmental, University, Environment	R
IQ13	**SEPTEMBER, Paperback, Saudi**	
44	ISBN, December, Press, Representation, Theory, Associative, Algebras, Assem, Skowronski, Cambridge, University, Dimensional, DynamicsKaren, Brucks, Henk	I
45	Arabia, East, Middle, Books, Hardcover, Editions, Long, Kingdom, Explore, General, Press, Products, Bestsellers, ASIN, Shipping	I
46	Author, ISBN, Publish, Description, Hardcover, Publishing, Customer, Found, August, Horror, October, Books, Editor, July, America	I

Table 3 identifies the novel clusters for this experiment with the labels for each cluster. There is a strong overall context of terrorism for most of the clusters obtained. However, the information retrieved covers a broad array of terrorism related topics that are outside the immediate domain of *"al qaeda" "bin Laden"*. For instance, a search on the top 5 words for cluster 31 *Enron, Afghanistan, UNOCAL, Cheney, India* retrieves novel web pages about the *Enron–Cheney–Taliban* connection that links Al Qaeda terrorism with the Enron corporate scandal. Taking another instance, the clusters derived from the intermediate query, *Terrorism, Osama, August (IQ10)*, detail the various terrorist groups with strong ties to Al Qaeda; how and why they originated; and the locations where they are active.

The results obtained from this experiment cover a broad range of useful information, such as

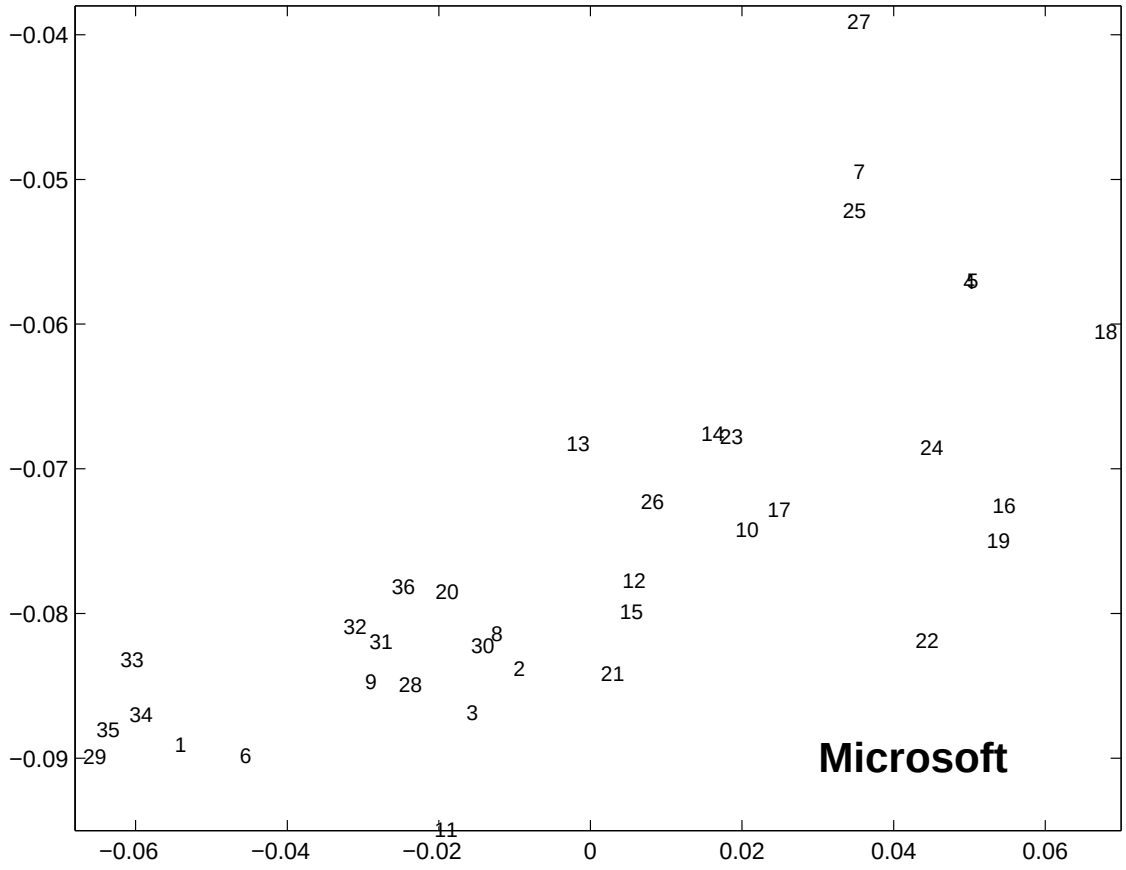

Figure 8: Novel Information Clusters

information about the *Iraq war-motive and implications*; *Al Qaeda* and the *Macedonian insurgency*; *Taliban* and its impact on Afghanistan; the *Carlyle group* and its links to the *bin Laden family*; *oil and terrorism*; *intelligence* and *Homeland Security*; *Islamic Jihad* (headed by bin Laden's deputy), among others. Most of the clusters obtained possess a strong relevance with terrorism and are outside the immediate scope of the original query.

The results for this experiment have been validated with the help of a subject expert. Table 4 represents the statistics associated with the validation process. A majority of the results have been classified as novel and very few as irrelevant.

Figure 9 is the two-dimensional plot obtained using SVD on the novel clusters and truncating. The groups of novel clusters show good coherence, for example the novel clusters pertaining to *Iraq Bush Saddam*. The plot shows a dense concentration of clusters which indicates that the information retrieved shared a broad underlying context, that is terrorism.

4.4 Results for the Query: Case Grammar

The goal for this experiment is to discover novel information relative to the query *Case Grammar*. The search for novel information is performed within the complete Web. The results have been rated by a subject expert according to the categories described in Section 3.4.2.

Table 5: Case Grammar Cluster Labels

IQ1	**Linguistics, University, Press**	
1	Oxford, Cambridge, Music, English, Science, Browse, Biography, Social, Dictionary, American, ISBN, National, Education, Humanities, Sciences	I
2	Series, Cambridge, Mathematics, Science, Sciences, Edinburgh, RESEARCH, Analysis, Applied, TECHNOLOGY, Humanities, Biology, Economic, TEACHING, Building	I
3	English, Oxford, Business, Classroom, Interactive, Headway, Global, Elementary, Policy, Legal, Rights, Reserved	I
IQ2	**Information, Theory, Learning**	
4	Inference, MacKay, David, Draft, Research, Cavendish, Laboratory, September, Gallager, Africa, Download, Publications, Barnes, Bayes	I
5	Education, Resources, Montessori, Childhood, Teachers, Development, Funderstanding, Concrete, Early, International, Library, Bernice, McCarthy, System, Left	I
6	Service, Women, Resources, MBTI, Community, University, Education, Kolb, Howard, Jung, MLnet, College, Research, Gender, Gardner	I

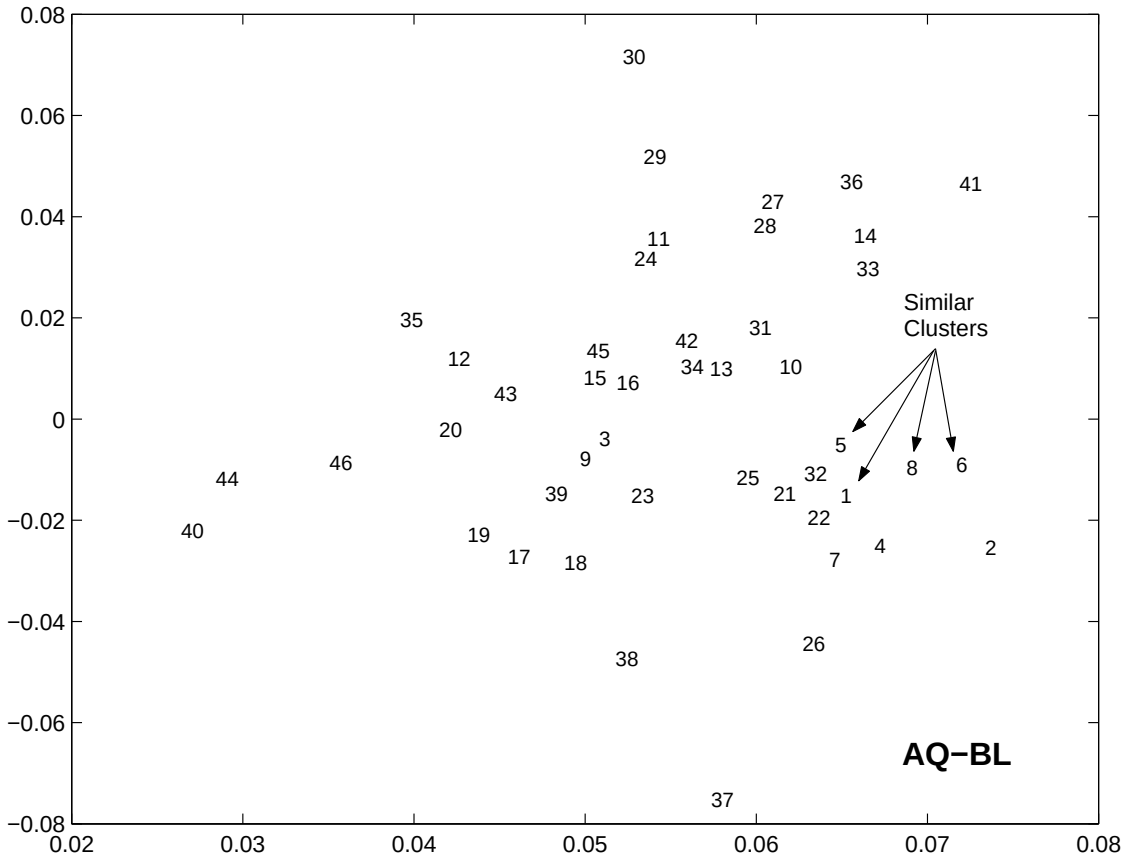

Figure 9: Novel Information Clusters

Table 5: Case Grammar Cluster Labels (continued)

7	London, Rogers, University, Press, Piaget, Routledge, Cambridge, Ramsden, Paul, Merriam, Caffarella, York, Skinner, Education, Wenger	I
8	David, MacKay, Inference, Algorithms, Lecture, Data, Noisy, Channel, Coding, Britain, Summary, Claude, Shannon	I
9	Neural, Networks, David, MacKay, Recognition, Summary, Supervisions, Prerequisites, Software, Part, Physics, Minor, January, Cambridge, Europe	I
10	MacKay, Rowling, Inference, Latin, Algorithms, David, Harry, Potter, Humour, Barnes, Noble, Welsh, Comparison, Corrections, Philosopher	I
11	Digraphs, Monographs, Algorithms, Jensen, Gregory, Gutin, Pages, Springer, Mathematics, ISBN, Bang, Amazon, Research, IMADA	I
IQ3	**English, Linguistics, Baylor**	
12	Anglo, Saxon, Indo, Beowulf, Languages, Middle, Entertainment, England, Yoda, Information, Georgetown, University, Resources, Lord, Literature	I
13	Indo, Germanic, Social-Sciences, Natural, Languages, Directory, Google, Science, Social, Sciences, Jobs, Press, Submit, Site, Open	I
14	Eastern, Michigan, University, Semantics, Psycholinguistics, Sociolinguistics, Typology, Program, Pray, Harrold, LINGUIST, Vernica, Grondona, Discourse, Analysis	I
15	University, Literature, Research, Professor, Interests, Lecturer, College, Spanish, Associate, Intermediate, Texas, Waco, Shanghai, Butler, Latin	I
16	LINGUIST, University, YORK, Department, Leeds, Lectureship, Applied, LECTURER, Steve, Harlow, LANGUAGE, SCIENCE, Mail, Personnel, French	I
17	University, Spanish, Literature, Twentieth, Texas, Latin, Language, Austin, Hispanic, Research, Interests, American, Graduate, Applied, Information	I
IQ4	**Lexical, Semantics, LINGUIST**	
18	Encyclopedia, Fablis, Linguistics, America, Doctorate, Languages, Computer, Wikipedia, Weather, Technology, Documentation, Word, Working	I
19	Chomsky, Speech, Rymer, Noam, Linguistics, English, Encyclopedia, Description, Research, Inter, Representation, Narrower, Saussure, Russ, Talmy	R
20	Baker, Wiese, Cognitive, Oracle, English, Publisher, Categories, University, Hardback, ISBN, Pages, Price, Science, Human, Error	PR
21	Greek, University, Linguistics, Natural, Language, Georgetown, Coptic, Siobhan, Mail, Prof, Department, Asst, Encyclopedia, Chapman, Directory	R
22	Language, Speech, Natural, Grammars, Connectionist, Phrases, Processing, Computational, Symbolic, Machine, Information, University, Research, Parsing	R

Table 5: Case Grammar Cluster Labels (continued)

23	Queries, John, Coleman, Michael, Nerbonne, Bruce, Stephen, Spackman, Mail, Compositional, Larry, Hutchinson, Nevin, Frege, Wlodek	I
24	pragmatics, Cognitive, Linguistics, Encyclopedia, Computer, Biographies, Fields, Encyclopedia4U, External, Documentation, Wikipedia, Lists, Countries	R
IQ5	**LINGUIST, Genitives, Unaccusatives**	
25	Unaccusative, English, Linguistics, University, Latin, Oshita, Peter, Query, Directory, Editor, Hiroyuki, Baltic, Dear, Matched	I
26	Linguistics, Semantics, Jobs, Ling, Confs, Diss, Syntax, USA12, Applied, Lang, Language, English, General, University	R
27	NELS, English, UMOP, China, University, Republic, Uighur, Turkic, East, Turkestan, Greek, Amherst, GLSA, Bible, Asia	I
28	Semantics, Phonology, University, Linguistics, Confs, Luke, Conference, Jobs, Syntax, Language, Disc, Hong, Ling, Kong, Workshop	R
29	Suffix-S, Affix, Morphology, Suffix-en1, Suffix-n, Suffix-e2, Suffix-e1, Suffix-es, Tutor, English, Left, DATR, Zwicky, Suffix-um, Suffix-er	R
30	Linguistics, English, Maxwell, Peter, Daniels, Ebonics, Fidelholtz, James, Sampson, Geoffrey, Steven, Schaufele, Homepage, Marilyn, Silva	I
31	Punctuation, Disc, Mail, Agreement, Hall, Bilge, Lexitron, University, Carl, Mills, Sandra, Summary, Robinson, Manchester, Mark	I
IQ6	**LINGUIST, Translation, Brown**	
32	PHONOLOGICAL, Phonetics, Anthropology, University, English, Interactive, THEORY, Linguistics, Department, Languages, Mail, Spring, International, Australia, POSITION	I
33	Chomsky, English, Linguistics, University, Language, Computational, Jobs, Auden, Machine, Proceedings, American, Research, John, Software, International	I
34	Languages, English, Linguistics, Human, Ethnologue, Evolution, Internet, Language, Science, Organized, Humanities, Resources, American, Foreign, Arts	I
35	Cognitive, English, Library, University, Sciences, Linguistics, Collection, Department, Science, Machine, Cinque, American, Speech, Computer, Pesetsky	I
36	Quine, English, LING, Stimulation, Quinean, Settling, Identification, Analytical, Wittgenstein, Philosophical	I
37	Quine, Gavagai, Radical, Rabbit, Stimulus, Rose, Analytic, Philosophy, Mark, Wilson, Sara, Leanne, Mastros, Augustinian, Twentieth	I
38	Endangered, Languages, Anthropological, Literature, Linguistics, Language, Press, Revitalization, Oral, Indigenous, Publishers, International, Publishing, Assoc	I
IQ7	**Identification, Model, Parser**	

Table 5: Case Grammar Cluster Labels (continued)

39	Word Research Tsai Chih Lexicon Lexical Analysis Summary dissertation University HTML Chinese Mental General Contents	R
IQ8	**Structure, Phrase, Bartlett**	
40	Quotations, Columbia, Bartleby, American, Roget, Thesaurus, English, Encyclopedia, Dictionary, Fiction, History, Usage, King, Wells, Stories	I
41	Psycholinguistic, Cognition, Psychology, Cognitive, Journal, Press, London, Processing, Cambridge, Research, American, University, Frazier, Foxtrot, Language	I
42	University, Sign, American, HPSG, Linguistics, Neidle, Aarons, Head, Kegl, Bahan, Phys, Chem, MacLaughlin, Conference, Levine	I
43	American, English, Quotations, Bartleby, Roget, Columbia, Verse, Dictionary, World, Oxford, Thesaurus, Encyclopedia, Usage, Poetry, King	I
44	HPSG, Computational, Linguistics, State, Resources, University, Interview, Homepage, Ohio, Site, Program, Head, Stanford, Systems	R
45	Journal, David, JOURNALS, Phycology, Paul, Cart, Author, Plant, ISSN, Marine, Freshwater, Contents, BLACKWELL, Morse, Genus	I
46	Checkpoint, WCAG, Extended, Core, Kynn, Checkpoints, Draft, August, Working, Accessibility, Suggestion, Guideline, Clear, Conformance, HTML	I
47	Kynn, WCAG, Accessibility, Loughborough, December, October, November, XHTML, Palmer, Sean, Mailing, Received, Public, HTML, Mail	I
48	Kynn, April, Thatcher, Alex, Mailing, Public, Technical, Developer, Liaison, Reef, North, America, Accessibility, Integrator	I
49	Kynn, Leonard, Kasday, WCAG, Exception, Essential, Purpose, General, October, December, XHTML, Checkpoint, Gilman, Single, Triple	I
IQ9	**Linguistics, Ethnologue, Publications**	
50	English, Languages, World, Language, Library, University, LOCATION, Seychelles, International, Dictionary, Internet, American, HTML, Institute, Green	I
51	Languages, Encyclopedia, World, Wikipedia, Language, Human, Documentation, Library, Argentina, University, Constructed, Dictionary, Sapir, Whorf, English	I
52	Harris, Journal, Jussi, Niemi, University, International, Language, Languages, Conference, Papers, Research, Abstract, Morris, Marja, Nenonen	I
53	Collection, StUB, General, Frankfurt, English, Linguistik, Bibliographie, Festschriften, Indo, Advice, Catalogues, Literature, Bibliotheken, Linguistischer, Literatur	I
54	LinguaLinks, Operator, Catalog, Languages, World, International, Field, BODY, Tools, Shopping, CAPTION, November, Library	I
55	Languages, Basque, English, Library, University, Databases, Electronic, Language, Resources, Dictionary, Natural, DeweyClass, Author, Location	I

Table 5: Case Grammar Cluster Labels (continued)

56	Languages, World, Catalog, Copyright, International, University, California, Weber, David, Huallaga, Quechua, Davidson, William, Western, Apache	I
57	Anthropology, Translation, Ethnography, Sociolinguistics, International, Contents, Place, Museum, Khmer, Journals, Institute, Journal, University, Textlinguistics, Texas	I
IQ10	**English, Transformational, Linguistics**	
58	Syntax, Structure, Theory, Morphology, Chomsky, Deep, Head, Oxford, London, Blackwell, Department, Chicago, Historical, Arnold, Professor	R
59	John, University, Press, York, Cambridge, Amsterdam, Benjamins, Sandra, Thompson, Paul, Hopper, Oxford, Bybee, Joan, Frequency	I
60	Computational, LANGUAGE, Press, England, Facsimile, Scolar, Cloth, Menston, John, Proceedings, Simmons, Thomas, Communications, Wing, Austin	I
61	University, American, Cambridge, Chicago, Press, York, Oxford, London, Corpus, Language, Research, Dictionary, John, Library, Word	I
62	University, Generative, Speech, Features, Brigham, Young, Edition, INTRODUCTION, Usage, Language, Variation, Historical, Regional, Social, Purdue	I
IQ11	**Avgustinova, Slavic, Conference**	
63	Library, Koerner, Geography, Perspective, Abstracts, Sciences, Social, Restricted, Citation, Canadian, Current, Science, English, Russian, Information	I
64	Makedonian, Slavko, Makedonia, Florina, Makedonians, Greek, Slavs, Hellenic, Bulgarians, Skopia, Ouranio, Ethnicity, Toxo, MGSA, Greece	I
65	HPSG, Linguistics, LARISA, ZLATIC, University, Wechsler, Stephen, Agreement, Serbian, Chicago, Stanford, Berkeley, Proceedings, Constraint, Formal	R
66	HPSG, Linguistics, Programme, Workshop, Languages, Papers, August, November, Formalisms, University, International, Project, European, Registry, Sites	R
67	Oliva, Simov, HPSG, Proceedings, Kiril, Linguistics, Workshop, Proc, HPSG, University, Osenova, Petya, Zipped, Corpora, Computational	I
68	Agreement, University, Austin, Morphology, Phrase, BREAK, Tania, Approach, Texas, March, Gender, Korean, Partial, Feature, Anti	I
69	Linguistics, University, Russian, Association, American, Languages, Language, Education, Conrad, East, Russia, International, Deadline, European, Century	I
70	HPSG, Przepiorkowski, Stanford, Adam, Borsley, Robert, CSLI, Publications, Structure, Head, Phrase, Serbo, Croatian, Penn, Overview	I
71	University, Union, Yates, Location, Missouri, Columbia, Kansas, Panel, Chair, Russian, State, Moscow, March, Coffee, Russia	I

Table 6: Expert Validation Statistics

	Number of Clusters	%age
Relevant	15	17
Partly Relevant	1	1
Irrelevant	76	82
Unknown	0	0
Total	92	100

Table 5: Case Grammar Cluster Labels (continued)

IQ12	**Pike, Kenneth, Matrix**	
72	Ethnologue, Publications, Site, University, International, Michigan, Linguistics, Citation, BODY, Catalog, LinguaLinks, Tools, Shopping, CAPTION, November	R
73	Linguistics, Saussure, Harris, Emic, William, Thomas, ETICS, EMICS, Univ, Charles, American, Institute, International, Summer, English, Dallas	R
74	Newton, University, Slave, English, Linguistics, John, Release, Extent, Language, Description, Institute, Diaspora, Chicago, York, Africa	I
75	Evelyn, Miguel, Grande, Ethnologue, Publications, Rachel, Saint, Faust, Norma, Cocama, Mexico, Site, Lanalyse, Contrastive, Anlisis	I
76	Harris, Ethnologue, Anthropology, Marvin, Headland, Zhuanglin, Thomas, Publications, Citation, Newbury, Park, Sage, Tagmemics, Reviews	R

Table 5 identifies the novel clusters for this experiment with the labels for each cluster. There is a strong overall context of linguistics for most of the clusters obtained. The information retrieved covers a broad array of topics about linguistics and language theory that are outside the immediate domain of *Case Grammar*. For instance, a search on the top 5 words for cluster 92 (*Harris Ethnologue Anthropology Marvin Headland* retrieves web pages about *Linguistic Anthropology*, a topic that links language development with human evolution. Taking another instance, the clusters derived from the query, *Information Theory Learning*, contain information about inference theory, which is the theory of grammatical learning patterns.

The results obtained from this experiment cover a broad range of useful information, such as information about the *Inference Theory and Information Theory*; *Graph patterns in Linguistics*; *Syntax and Semantic theories of grammar*; *Speech pattern and Phonetics*; *Web Content Accessibility group*; *Head–Driven Phrase Structure Grammar or HPSG* and *Transformation Grammars – Deep Structure and Surface Structure*, among others. Most of the clusters obtained possess a strong relevance with grammatical study of languages.

The results for this experiment have been validated with the help of a subject expert. Table 6 represents the statistics associated with the validation process. Only 17% of the total results were classified as relevant.

Further, as mentioned in Section 3.3.3, the words representing the cluster labels are not necessarily ontologically coherent, even though the underlying cluster might be. In other words, a set of nouns in Table 5 may not appear to be naturally cohesive (or a well-defined topic representa-

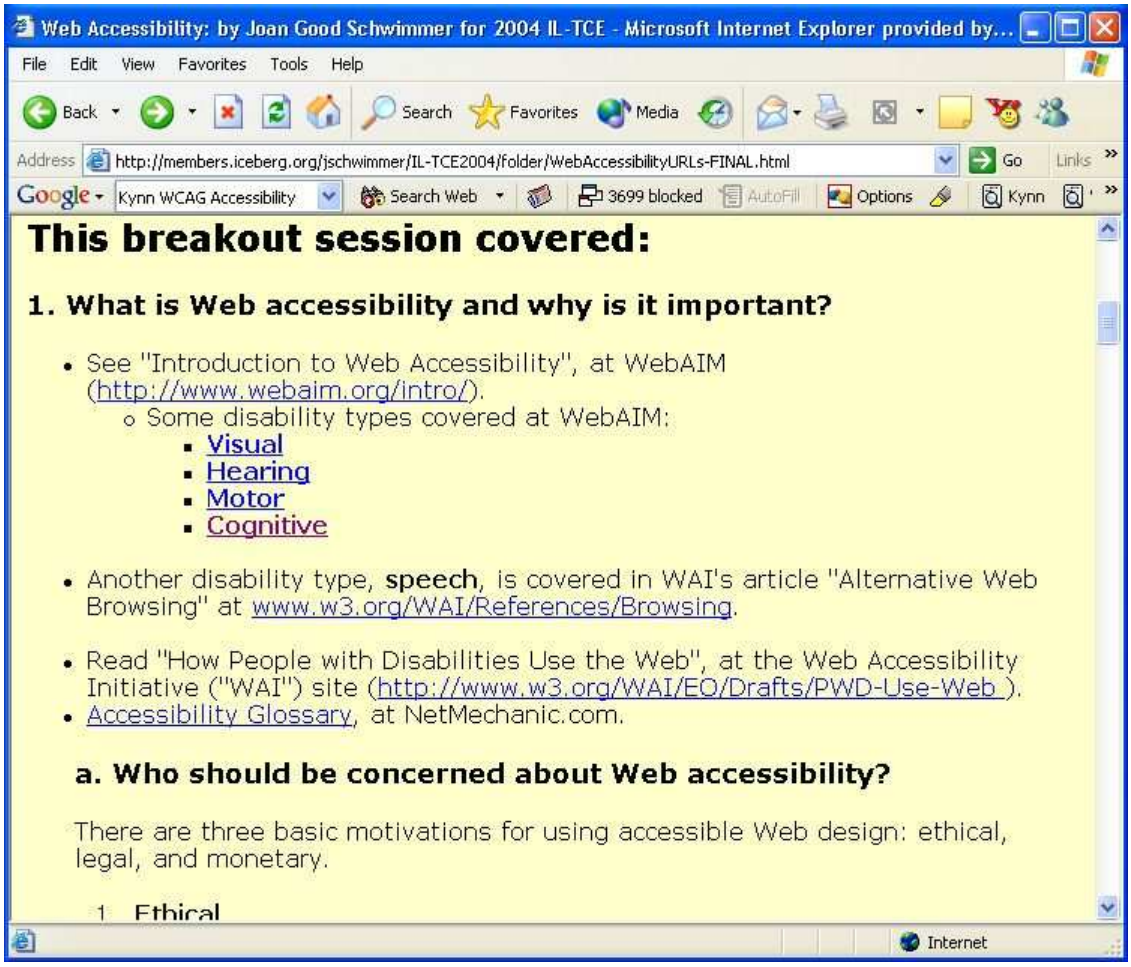

Figure 10: Web page retrieved using the query: *Kynn WCAG Accessibility*

tion) to an individual. However, a search on these words is likely to retrieve useful, and related, information. For instance, cluster 47 is classified as irrelevant and does not appear to be a cohesive topic representation. However, a search on the top 3 words (*Kynn WCAG accessibility*) reveals information about *Web Content Accessibility*, which includes linguistic features to make web pages comprehensible. Figure 10 shows a web page about Web accessibility initiatives that was retrieved using this query.

Figure 11 is the two–dimensional plot obtained after SVD on the novel clusters obtained and truncating to two dimensions. The groups of novel clusters show good coherence, for example the novel clusters pertaining to *Information Theory Learning*. There is an unusually dense concentration of clusters in the plot. It can also be noticed that the clusters pertaining to *Ethnologue Publications* (the top-right of the plot) are isolated from this concentration.

4.5 Results for the Query: Multiple Sclerosis

The goal for this experiment is to discover novel information relative to the query *Multiple Sclerosis*. The search for novel information is again performed within the complete World Wide Web. The results have been rated by a subject expert according to the categories described in Section 3.4.2.

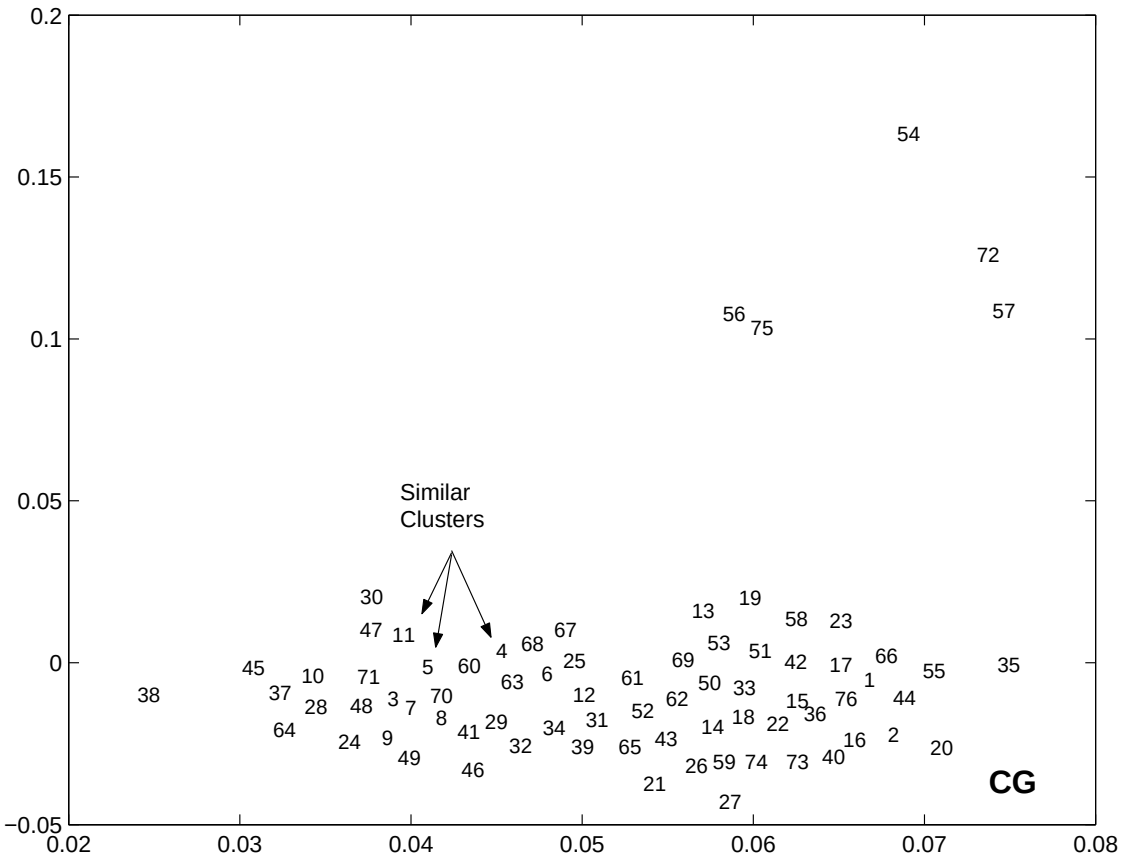

Figure 11: Novel Information Clusters

Table 7: Multiple Sclerosis Cluster Labels

IQ1	**National, Cancer, Myeloma**	
1	Treatment, Hodgkins, Institute, Lymphoma, Hodgkin, Clinical, Site, Chemo, Lung, Service, Multiple, Prostate, Dictionary, Publications, Information	PR
2	Leukemia, Mayo, Clinic, Multiple, Research, Scottsdale, Foundation, Institute, Disease, Plasma, Hodgkin, Information, Lymphoma, Cell, Plasmacytoma	PR
3	Myelodysplastic, Treatment, Syndrome, Institute, Description, Syndromes, Chemo, Site, Stage, Explanation, Option, Overview, Novo, Treated, Summary	R
4	Lymphoma, Leukemia, American, Treatment, Institute, Hodgkin, Health, Childhood, Information, Blood, Clinical, Marrow, Research, Breast, Statistics	R
5	MULTIPLE, Research, Institute, Foundation, ASSOCIATION, Medical, Arkansas, Clinic, SCCA, FHCRC, Anderson, Lipper, University, McCarty, Gammopathy	!

Table 7: Multiple Sclerosis Cluster Labels (continued)

6	Remission, Cleveland, Clinic, Family, Health, Information, Site, Taussig, MlSH, Institutes, Cured, Contact, Institute, Policy, HONcode	!
IQ2	**Research, National, Institute**	
7	Health, Safety, Mining, NIOSH, Injury, Materials, State, Penn, Occupational, Emergency, Prevention, Training, University, Relationship, Enhancement	I
8	Health, Dental, Information, Cancer, ORAL, Clinical, Resources, Grants, Craniofacial, Site, Environmental, Group, NCFPR, Medicine, FUNDING	
9	Health, Clinical, Gordon, Oral, Cancer, Eastman, Chelation, Trials, General, NICHD, Solicitation, EDTA, Apply, NIST, Deadline	!
10	Centre, USADr, Nutrition, University, Human, Health, Genome, Dental, Singapore, UKDr, Oral, Bethesda, Medicine, Washington, Cambridge	R
11	Cancer, Site, Treatment, Funding, Clinical, Health, Cancers, Genetics, Trials, Publications, Study, Prevention, Information, Resources, Education	R
12	University, Cancer, Boston, Molecular, Fellowship, School, Hospital, Medicine, York, Identification, California, CLINIC, Postdoctoral, Jolla, Mechanisms	R
IQ3	**Health, Stroke, Diseases**	
13	Alzheimer, Association, Heart, National, American, Multiple, Dementia, Institute, Pages, Information, Women, Research, Education, Blood	PR
14	Americans, NINDS, Information, American, National, Disorders, Whites, Institute, African, Women, Rehabilitation, Neurological, Indians, Asians, Pacific	R
15	National, Alzheimer, Disorders, Neurological, Institute, Association, American, Heart, University, Medical, Information, Research, PubMed, Parkinson, Neurology	R
16	Heart, National, Alzheimer, Disorders, Neurological, Institute, Association, NINDS, American, Jakob, Lung, Institutes, McIlroy, Creutzfeldt, Policy	R
17	Association, American, Lung, Camp, Heart, Abnaki, Research, Disease, AIDS, Genital, Molluscum, Contagiosum, Scabies, Human, World	!
18	Classification, NCHS, International, Disability, Functioning, Revision, Rich, Classifications, Clinical, Modification, Maintenance, Committee, NACC, Policy, Accessibility	R
19	Disorders, York, Internet, Information, Merck, Library, Medicine, University, NOAH, National, Medical, Consensus, Control, Hardin, TMJD	R
20	Heart, Alzheimer, Disorders, Neurological, Neurology, Coronary, Medical, WebMD, EUROPEAN, View, Word, April, Monday, Information, Research	I
21	Information, Asthma, Lung, NHLBI, Facts, National, Pulmonary, Chronic, Publications, COPD, Fact, Disease, School, Sheets, Emphysema	!

Table 7: Multiple Sclerosis Cluster Labels (continued)

22	National, Information, Institute, Treatment, Research, Rehabilitation, Association, Aging, NINDS, Institutes, Control, Prevention, Bethesda, Neurological, American	R
IQ4	**NINDS, National, Disorders**	
23	Parkinson, Stroke, Disease, Institute, Neurological, Health, Syndrome, NORD, Research, Institutes, Foundation, Information, Barr, Developmental, Guillain	R
24	BrainNet, Health, Internet, Agency, Pharmaceutical, Company, Mail, Healthcare, Pharma, NeuroNews, Register, Contact, Mission, Decade, Publications	R
25	Diseases, Guide, Orphan, Institute, Neurological, Stroke, Syndrome, SyndromeArticle, SyndromeInformation, System, Articles, Asperger, Huntington, Disease, Parkinson	R
26	Cerebral, Palsy, Health, Neurological, Institute, Stroke, Information, Diseases, Syndrome, NORD, Pregnancy, Alzheimer, Research, Describes, Institutes	R
27	Stroke, Neurological, Institute, Acute, Institutes, Brain, Street, Health, Preventing, Information, Avenue, Research, York, Browse, Rehabilitation	I
28	Syndrome, Apert, Iron, Institute, Marfan, Hemochromatosis, Foundation, Guide, Information, House, Health, Craniofacial, Site, Statement, Aran	!
29	Genetic, Opitz, Site, Disease, Translation, Health, Communication, Medical, Genetics, English, University, NCNCD, Neurogenic, Miller, Weber	?
30	NIOSH, Ergonomics, Musculoskeletal, Redirector, Muskuloskeletal, Topic, Skin, Centre	I
31	Syndrome, Anemia, NORD, Disease, Deficiency, Database, Hemolytic, Dysplasia, Acidemia, Hereditary, Diseases, Type, Organizations, Congenital, Primary	R
32	Muscular, Dystrophy, Duchenne, myotonic, Stroke, Institute, Neurological, Browse, Institutes, Publications, Emery, Dreifuss, Information, North, Dystrophy	R
IQ5	**National, Health, International**	
33	Social, AIDS, Travel, Information, Emergency, Security, Fund, Development, Africa, Global, State, Research, World, Insurance, Prevention	R
34	SCIEH, Scottish, Centre, Environmental, Travel, Medicine, Division, Scotland, Receive, Journey, Advice, Tips, British, Embassies, IVTelecom	R
35	Golf, South, Africa, Social, Marie, Stopes, Travel, Quotes, Specials, Vacation, Finder, Policy, Contact, Ministry, Affairs	I
36	Yellow, Canada, SARS, Capital, Altitude, Ipas, Scouts, Pages, Department, Islands, Travel, United, World, Africa, Statistics	I
37	Music, General, IERHB, Yellow, American, University, Services, World, Pages, Bradley, Public, United, Insurance, Group, Africa	I

Table 7: Multiple Sclerosis Cluster Labels (continued)

38	NIAID, January, Africa, DEPARTMENT, South, World, Media, SERVICES, Kenya, Weather, HUMAN, Foundation, Release, Resistance, Site	PR
39	Calendar, Yellow, Sheet, Institutes	I
40	Internet, Preferences, Options, Browsealoud, Accessibility, Fonts, Explorer, Netscape, Tools, General, Change, Tick, Edit, Font, Zoom	I
41	Island, Puerto, City, Santa, Port, Park, Beach, North, Cape, South, Lake, Marina, Saint, America, Salvador	I
IQ6	**NINDS, Lateral, Amyotrophic**	
42	Primary, Symptoms, Diseases, DISCLAIMERS, Misdiagnosis, Prevalence, Lists, WrongDiagnosis, Information, List, Difficulty, Progressive, Summary, Class	R
43	Gehrig, Neurology, Association, Information, Disease, Medications, Foundation, Anti, Motor, World, Federation, Practice, Ohio, Oxidants	R
44	Research, Federation, Neurology, Alzheimer, World, Group, Motor, Neuron, Diseases, Information, Committee, Forum, Neurological, Stroke	R
45	Neurological, Gehrig, Disorders, Diseases, Health, Disease, Information, National, Doug, Institute, Stroke, Motor, Neuron, PALS	R
46	Diseases, Guide, Orphan, National, Disorders, Institute, Neurological, Stroke, Disease, Neurol, Resources, Research, Medical, Information	R
47	Gehrig, National, Federation, Disease, World, Information, Disorders, Depression, Medicine, Survival, Guide, Advocacy, Advanced, Neurology	R
48	Symptoms, DISCLAIMERS, Diseases, Summary, ADVERTISEMENT, Incidence, Prognosis, Prevalence, Lists, Rate, Information, Misdiagnosis, Contents, Americans	R
49	Alliance, International, Associations, Site, Study, Association, Experts, Allied, Professionals, Forum, Directory, Board, Clicking, internet, Stem	R
50	Disease, Treatment, Gehrig, Information, Health, Drug, Resources, National, Association, Diseases, Neurology, Yahoo, Guide, Cancer	R
51	Health, Information, Disorders, Disease, Neurological, Diseases, Gehrig, Medical, Directory, PALS, Guide, Internet, Google, Associations	R
IQ7	**Tony, Winter, Tenen**	
52	Fruits, Veggies, Rich, Phytochemicals, Blueberries, Broccoli, Cherries, Citrus, Collard, Garlic, Kale, Onions, Pink, Spinach, Strawberries	R
53	Abstract, Cell, Biol, Cancer, Science, Methylation, Chem, Receptor, Department, Retinoic, Pelicci, Italy, Hypermethylation, Leukemia, Protein	R
54	Math, Fall, Spring, Algebra, Class, Calculus, Linear, Iarrobino, Info, Francisco, Hilbert, Information, Graphing, Calculator, Section	I
55	Institute, University, Research, France, Myelopoiesis, Boston, COFFEE, BREAK, Saint, Louis, leukaemia, Paris, Sunday, Harvard, Tuesday	?
56	Division1st, Johnny, Righteously, Toad, Open, League, Curtis, Division, Group, Warning, John, Weather, Rance, 11st, Craig	I
IQ8	**General, Neurology, BrainTalk**	

Table 7: Multiple Sclerosis Cluster Labels (continued)

57	Neurosurgery, Medscape, Stroke, MEDLINE, Epilepsy, Neuroimaging, Parkinson, Disease, Syndrome, Seizures, Abstracts, Case, Challenge, Multiple, Sclerosis	R
58	Cochrane, Epilepsy, Review, Library, Group, Register, January, Effectiveness, Software, Abstract, Objectives, Bradley, Controlled, Trials, Lindsay	R
59	January, Ring, Communities, Epilepsy, Spinal, Support, WebRing, Chronic, Site, Parkinson, Chiari, Injury, Disorders, Multiple, Autism	R
60	Neurological, Communities, Disorders, Health, List, Support, Medical, Resources, Family, Cord, Lester, Sites, Policy, American, Internet	R
61	info, Disease, Information, MEDICAL, Support, Resources, Parenting, Service, Partners, Health, Learning, Mothers, Parent, Disabilities, Huntington	R
62	Hospital, Syndrome, Disorders, Disease, Medical, Hornsby, Show, Health, Lester, Medicine, Multiple, Support, Parkinson, Community, Interns	R
63	Support, Communities, Directory, Respiratory, Asthma, Disease, Wednesday, Organizations, Pediatric, Pages, Groups, Treatments, Rings, Diseases, Listing	!
64	WebRing, Ring, Communities, Site, RingSurf, Directory, Password, Random, Contact, Owner, Submit, EMail, Address, Code, Membership	I
65	John, Logged, Communities, Lester, Support, Category, Neurological, Disease, List, Topic, Multiple, Health, Forum, Disorders	R
66	Parkinson, Disease, National, Foundation, Research, Neurological, Institute, Disorders, Information, Stroke, Canada, Association, Pages, American, Drug	R
67	Disorders, Syndrome, Disease, Multiple, Support, Disorder, Injury, Spinal, Category, Communities, Medical, Child, Neurological, Chronic	R
68	Cochrane, Epilepsy, Update, Software, Library, Review, Lindsay, Group, Register, Bradley, Effectiveness, Oxford, Controlled, Trials, BABSTRACTA	?
69	Disorders, Syndrome, Injury, Disease, Chronic, Forest, Spinal, Communities, Neurological, Rehabilitation, Support, Cord, Stroke, Disability, Medical	R
70	Disorders, Syndrome, Disease, Show, Support, Multiple, Spinal, Disorder, Neurological, Moderator, Injury, Medical, Child, Chronic	R
71	Massachusetts, Hospital, Harvard, Boston, Webforums, Communities, Clinic, Surgery, Department, School, Information, Tumor, Medical, University, Radiology	R
72	Fibromyalgia, Syndrome, Chronic, Fatigue, Association, America, Immune, Dysfunction, Arthritis, Treatment, Research, Univ, Washington, Diagnosis, Foundation	R
73	Services, Rush, Clinical, Neurological, Myasthenia, Chicago, Gravis, Alzheimer, University, Fragile, Syndrome, Vascular, Neurosurgery, Program, Neurologists	R

Table 7: Multiple Sclerosis Cluster Labels (continued)

74	List, Syndrome, Chronic, Fatigue, Health, Clinical, Library, Sciences, Resources, INFO, Consult, Medicine, Diseases, Trials	R
75	Jaymie, Disorders, Communities, Ulnar, Policy, Dictionary, Moved, REGISTER, Nerve, CHRONIC, LEFT, PAIN, Internet, Massachusetts, John	?
IQ9	**Migraine, Neurological, OpenForm**	
76	MEDLINE, Abstract, VGCC, Channels, CALCIUM, Free, PubMed, Drosophila, Order, NERVE, Neurosci, Infotrieve, Citation, VGCCs, KAWASAKI	R
77	Medline, Abstract, Free, Biol, Cell, CrossRef, Camilli, Chem, Acad, Proc, Natl, Wenk, Science, Curr, Nature	R
78	CaV2, Abstract, PubMed, Free, Biol, Medline, Order, Infotrieve, Neurosci, Cell, Chem, CrossRef, Margolis, Neurocan, Binding	?
79	Abstract, Neurosci, Channel, Biol, Chem, Channels, PNAS, Syntaxin, Voltage, Catterall, Interaction, January, April, gated, SNAP	?
80	Nasel, Effects, Fragrance, Americans, Fragranced, Fragrances, Modern, Materials, Citral, Headaches, Natural, Chem, Senses, Samec, Schindler	I
81	Logged, Military, Flying, Bonnie, USARegistered, General, Head, Neurology, Policy, Disorders, Topic, Class, Force, Greg, Support	R
IQ10	**Health, National, Information**	
82	Breastfeeding, Campaign, NWHIC, Office, Women, Breastfed, Contact, Human, Department, Awareness, PSAs, Advertising, Helpline, CDPs, Council	!
83	Hormone, Women, Internet, Address, Therapy, Initiative, American, Postmenopausal, NWHIC, Heart, Institute, NHLBI, Institutes, July, Progestin	R
84	Institute, Therapy, Hormone, Study, Women, Aging, Heart, Estrogen, Initiative, Menopause, Postmenopausal, Blood, Lung, Cancer, Medicine	!

Table 7 identifies the novel clusters for this experiment with the labels for each cluster. Clearly, most of the novel information retrieved concerns neurological diseases, along with the associated research and rehabilitation sectors. For most of the clusters, a search using the top few nouns reveals useful novel information relevant to *Multiple Sclerosis*. For instance, a search on the top 5 words for cluster 32 (*Muscular Dystrophy Duchenne myotonic Stroke*) retrieves novel web pages about *Muscular Dystrophy*, a disorder that affects the muscular system and shares many common symptoms with *Multiple Sclerosis*. Taking another instance, the clusters derived from the query, *General Neurology BrainTalk*, contain information about communities that support and rehabilitate patients with neurological disorders.

The results obtained from this experiment cover a broad range of useful information, such as information about the *Multiple Myeloma* and other cancerous disorders that affect the nervous system; *Neurological Disorders* such as *Cerebral Palsy*, *Hydrocephalus*, *Creutzfeldt–Jakob disease* and *Tourette Syndrome*; *Alzheimer's* and *Parkinson's* disease; *BrainTalk communities* for online support of neurological patients; *Genetic disorders*, among others. Most of the clusters obtained possess a strong relevance to the health and rehabilitation domains.

Table 8: Expert Validation Statistics

	Number of Clusters	%age
Relevant	50	60
Partly Relevant	4	5
Irrelevant	14	16.5
Not–Thought–Of	10	12
Unknown	6	6.5
Total	92	100

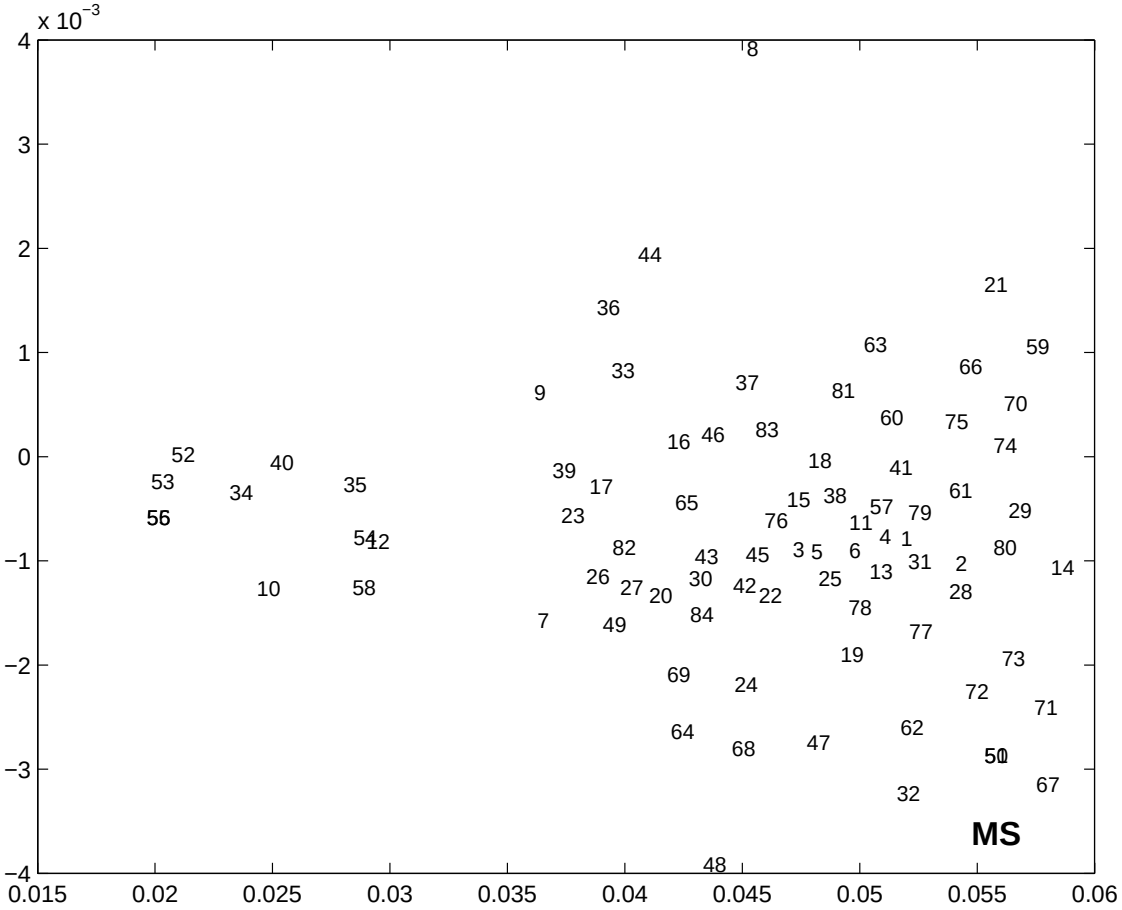

Figure 12: Novel Information Clusters

The results for this experiment have been validated with the help of a subject expert. Table 8 represents the statistics associated with the validation process. Only 16% of the total results were classified as irrelevant. Most of the results represented information about various neurological disorders, support communities and health research.

Figure 12 is the two–dimensional plot obtained after SVD on the novel clusters obtained and truncating to two dimensions. The groups of novel clusters overlap significantly and no particular group of clusters is completely distinct. Thus it presents a picture of related information.

4.6 Results for the Query: Multiple Sclerosis (.uk)

The goal for this experiment is to discover novel information relative to the query *Multiple Sclerosis* in the *.uk* domain. The results can then be compared to the general Multiple Sclerosis experiment (see Section 4.5). This gives interesting insights into the organization of information inside a specific domain of the Web. The results in tabular and graphical form are presented below. As always, the tabular form comprises each intermediate query (**IQ**) along with the 15–noun labels for the associated novel clusters. The results have been rated by a subject expert according to the categories described in Section 3.4.2.

Table 9: Multiple Sclerosis (.uk) Cluster Labels

IQ1	**Health, Crisis, Action**	
1	NORTHERN, IRELAND, Hospital, Ulster, Belfast, Hendron, Drumcree, World, Education, Entertainment, Depth, Front, Altnagelvin, Politics, Business	I
2	Nigeria, IRELAND, Business, World, NORTHERN, Africa, Malawi, Point, Entertainment, Front, Depth, Politics, Education, Abacha, Drumcree	I
3	STIs, Sexual, Education, Chlamydia, Ingham, World, National, Wales, Business, Politics, Science, Nature, Technology, Entertainment, Magazine	!
4	STIs, Sexual, Plan, Teenage, Pregnancy, July, National, Adobe, undiagnosed, Reader, Serv, Unit	!
5	STIs, Guardian, Professor, Adler, Britain, July, Government, Syphilis, Unlimited, Mail, Arts, SEXplained, Sexual, Review, Sign	!
6	Commission, Forum, ECPP, European, Public, Information, Core, NGOs, Rights, Malawi, General, Declaration, Human, Conferences, Patient	I
7	Malawi, Southern, World, Francis, Bank, Government, Economic, Justice, Network, International, Monetary, Fund, Africa, Winter, Word	I
IQ2	**Health, Illnesses, National**	
8	Depression, Alzheimer, Disease, Heart, Syndrome, Ailments, Epilepsy, Doctor, GUIDE, Message, Board, Awareness, Allergies, Arthritis, Asthma	R
9	Spinal, Disabilities, Body, GOSH, Bifida, Autism, Lupus, LIFE, Sneak, Games, HOSPITAL, Dear, Dictionary, Treatment	R
10	Hydronephrosis, Addictions, Allergies, Arthritis, Asthma, GUIDE, Blocking, Preventing, Doctor, Message, Board, QUICK, LiveA, Ailments, DEPTH	R
11	Diabetes, Kidney, Organisations, Disease, Hospital, Guide, EASDEC, Road, London, Lifestyle, Department, Moorfields, Pituitary, Foundation, Insipidus	R
12	Impotence, Osteoporosis, Campaigns, Awareness, Doctor, Message, Board, Samaritan, Homepage, Lifestyle, Programmes, Radio, Relationships, Interactive, Area	R
13	Mental, Alliance, Depression, Fellowship, Association, General, England, Wales, MACA, Point, Rethink, YoungMinds, UNISON, Richmond, Crisis	I

Table 9: Multiple Sclerosis (.uk) Cluster Labels (continued)

14	Schizophrenia, Information, Mental, Research, Disorders, South, Support, Nash, United, Project, Articles, Korea, Clinical	I
IQ3	**Education, Conductive, National**	
15	Health, Wales, Learning, Foundation, Government, Association, England, Information, Advice, BETT, Institute, United, Assembly, Becta, Teachers	I
16	Institute, Grand, Central, Contact, Foundation, Guide, Canada, Germany, Zealand, North, America, Internet, Close, Database, Health	I
17	Institute, Charity, Library, Guide, Russell, Birmingham, Email, Information, Resources, Foundation, Nominate, Database, Health, Open, Close	I
18	Parkinsons, Disease, London, Press, Brown, M2Aust, Film, University, Library, Journal, Research, British, Horvth, Mikula, Toth	R
19	Jacqueline, Institute, Conductors, Foundation, Guestbook, Hungary, Birmingham, Chronicle, Community, Inclusion, Centre, Market, Place, Connections, Library	I
20	Cerebral, Palsy, Patient, Centre, Hearing, Hemiplegia, Support, Information, Motor, Forward, Medicine, Medline, Leaflets	R
21	Dyspraxia, Hantsweb, Institute, Hampshire, County, Council, Cousin, Homepage, Area, Organisation, Gillian, Maguire, Mail, Address, Cannon	R
22	Association, Sick, Patient, NAESC, Special, Publications, Department, Office, Hospital, Information, HouseHerald, WayPegasus, England, Wales, LEAs	I
IQ4	**Centres, National, Therapy**	
23	BTEC, Science, Edexcel, Applied, Materials, Technology, Sciences, Diploma, Design, Services, Level, Foundation, Beauty, International, Qualifications	I
24	Beauty, MAPPs, Training, Sector, Authority, Wales, Northern, Ireland, NTOs, Industry, Councils, Saturdays, NVQs, SVQs, Levels	I
25	Yoga, Dance, Training, Therapists, Association, Movement, ADMT, Diploma, DMTs, Opportunities, London, business, Cymru, Wales, Northern	I
26	Leisure, Sports, West, Sussex, Park, Surrey, Complex, Recreation, Stadium, East, Cheshire, Nottinghamshire, Berkshire, London, Midlands	I
27	Occupational, NHSU, Therapists, Speech, Health, Qualifications, Wales, Northern, Ireland, College, Professions, RCSLT, Council, GCSEs, Authority	R
28	Therapist, Health, NeLH, Speech, Internet, Portal, Research, Library, Clinical, Database, Public, Register, Frameworks, Libraries, Diseases	R
IQ5	**Research, CancerHelp, Cancer**	
29	Breast, Copyright, Information, Site, Glossary, Printer, Clinical, Trials, Donate, Access, Keys, Direct, Partners, Programme, Radiotherapy	I

Table 9: Multiple Sclerosis (.uk) Cluster Labels (continued)

30	Health, Author, DeweyClass, ResourceType, Location, Oncology, University, Internet, Resources, National, Information, Institute, Association, International, Cure	R
31	Health, Mesothelioma, Information, Centre, Medical, Breast, Free, Clinical, Institute, British, Staff, Macmillan, Foundation, Resources, Lung	PR
32	Oncology, Leukaemia, Resources, RMCS, IBST, SIMS, Medical, Cell, Zioupos, Gloucester, Turner, Tumours, Internet, Lovell	R
33	Birmingham, Clinical, KPMG, Claire, Annie, Board, Campaign, Trials, University, Nick, Institute, Information, James, Site, Hospital	PR
34	Donate, Newsletter, Support, Science, Runs, Race, International, Treks, Trek, SunSmart, Site, England, Wales, Lincoln	I
35	Health, Macmillan, Fund, Centre, BACUP, National, Institute, Internet, Relief, Vanderbilt, Gateway, Breast, Campaign, Cancerweb, Committee	?
36	Life, Race, Donate, websites, JavaScript, Direct, Debit, Runs, Bobby, Moore, Fund, International, Treks, Running, Stride	I
IQ6	**Christmas, Cards, Charity**	
37	Credit, AstraZeneca, Scottish, Trust, Research, Insurance, Cancer, Widows, Saver, National, Royal, Macmillan, Green, Foundation, Worcester	I
38	Visit, Macclesfield, Wilmslow, Trust, Research, Cancer, Cross, Manchester, Hospital, Arthritis, Blue, Centre, Moor, Hale, Campaign	I
39	Amnesty, International, Personalised, Trust, View, Card, Greetings, Collection, AIUKCT, Group, United, Kingdom, Rosebery, Avenue, London	I
40	Personalised, View, Tree, Photo, Front, Images, Decorations, Keyword, Card, Shop, Humorous, Show, Time, Service, Logo	I
41	Personalised, Company, View, Greetings, Card, Personalisation, Services, Generator, Royalties, Trading, Ordering	I
42	DMTs, Neurology, National, Library, Health, Information, Institute, Medical, ukMS, Internet, Teva, comMS, American	R
43	Neurology, Journal, Cochrane, Blumhardt, Psychiatry, Neurosurgery, Study, Sawle, Group, Health, Suppl, Parkinson, Edwards	R
44	Rebif, Betaseron, Avonex, Singer, Medscape, PRISMS, Neurology, Barry, Linomide, Serono, LTFU, Pfizer, Disability	R
45	Trust, British, Health, Genetics, Association, Cancer, Royal, College, Genetic, Head, Group, Services, Human, Public, Committee	R
46	Psychology, Research, Health, Osteoporosis, Bone, Department, Professor, Lecturer, CPsychol, Group, Senior, Royal, University, College, Tutor	R
IQ7	**Copaxone, EDSS, Trust**	
47	Consultant, Neurologist, Interferon, Beta, Research, MSRC, Health, Scheme, Service, Wolstenholme, Price, Availability, Clinical	R
48	Information, Betaseron, Hotlinks, Internet, Resources, Books, International, Disabled, Foundation, Holiday, RADAR, Biogen, Pharmaceuticals, Facts, Australia	R

Table 9: Multiple Sclerosis (.uk) Cluster Labels (continued)

49	Disability, Scale, Hospital, Neurological, GNDS, Information, Updates, Thomas, Exercises, LondonThe, activityMagnetic, Resonance, Imaging	R
50	Betaferon, Rebif, SPMS, Lancet, Interferon, Group, Neurology, RRMS, Study, Office, Belfast, Avonex, Interferons	R
IQ8	**Bandolier, Cannabis, Neurology**	
51	Journal, Hospital, Medical, Panel, Medicine, Archives, Royal, Revista, General, Surgery, Research, London, Road, Department, American	R
52	Oxbridge, Solutions, GPnotebook, Limited, Rights, America, Patient, Practice, North, Notebook, South, Africa, Asia, Australasia, Agreement	I
IQ9	**Bandolier, EDSS, Contact**	
53	Internet, Future, Problem, Site, Journal, Extra, Glossary	I
54	Study, JAMA, Colorado, Information, Comment, Baseline, November, February, Interventions, Brownie, Refrigerator, Eligible, Reduction, Reference, Gonzales	I
55	Internet, Journal, Site, Extra, Glossary, Aerobic, MIDAS, Stratified, Cannabis, Independence, Clinical, Reference, Medicine, Physical, Penninx	I
IQ10	**Health, Medical, Fertility**	
56	Discovery, Pregnancy, Chlamydia, General, Women, Its, Parenting, Teen, Personal, Channels, Images, Photodisc, Copyright, Communications, Data	I
57	Creutzfeldt, Jakob, World, Professor, Prusiner, England, Prion, University, Science, Stanley, Nature, Prions, Journal, Medicine, Gerstmann	?
58	Discovery, Hormonal, Mumps, General, Cancer, Affects, Lindane, Testicular, Sperm, Women, Pregnancy, Parenting, Teen, Personal, Channels	PR
59	Cycle, Buddies, General, Chit, Series, Discovery, Community, Friends, Infertility, December, October, YaBB, March, April, Media	I
60	World, Human, Fertilisation, Embryology, Northern, Ireland, Scotland, Wales, Business, Politics, Education, Science, Nature, Technology, Entertainment	I
61	Guide, Fitness, Sites, Women, Advice, Private, London, Natural, Personal, Injury, Partner, FertilityFertilityAbout, LinksFancy, Great, Directory	R
62	Guide, Sites, Fitness, Private, TheComment, Insurance, Personal, Life, Norwich, Union, Watchers, Advice, Healthcare, Affordable, CoverLet	PR
IQ11	**Priory, Hospital, School**	
63	House, Services, Grange, North, Rehabilitation, Consultant, Clinic, College, Grove, Ticehurst, Contact, Location, Farleigh, Hall, Education	I
64	Sturt, Addiction, Treatment, Program, Address, Lane, Walton, Hill, Surrey, KT20, Telephone, Description, Drug, Minnesota, Physical	I
65	Services, Education, Consultant, Working, General, Rehabilitation, ADHD, West, Legislation, Clinical, Accessing, Referral, Paths, Funding, Location	I

Table 9: Multiple Sclerosis (.uk) Cluster Labels (continued)

66	Nervosa, Programmes, Peter, Rowan, Eating, Patient, Anorexia, Disorders, Bulimia, Information, Unit, Description, Referral, Treatment, Request	R
67	Nervosa, Bulimia, Anorexia, Multi, Anorexic, Treatment, Rowan, Vomiting, Peter, Information, Topic, Description, Bulimic, Laxative, Major	R
68	Medical, George, Treatment, Georges, Nursing, Prize, National, Found, London, Tooting, Events, Hope, Intranet, Meningitis, AIDS	I
69	London, Kingdom, Roehampton, South, United, England, Information, Topic, Building, Strawberry, Richmond, Park, Central, West, Westminster	I
IQ12	**Health, World, Beta**	
70	Information, National, Research, Mill, Hill, Medical, University, Screening, England, Institute, Education, Journal, Thalassaemia, Newcastle, Lock	I
71	Thalassaemia, Haemoglobin, Information, Anaemia, Implications, Friday, Introduction, Mediterranean, Middle, East, Asia, Northern, Europe, Worldwide, Internet	R
72	Chancellor, University, Dundee, Vice, London, Langlands, Alan, Glasgow, Black, Designers, Design, James, Smith, Tuesday, Duncan	I
73	Depression, Encyclopaedia, Introduction, Topic, Symptoms, Diagnosis, Treatment, HealthSpace, Site, Direct, Complications, Prevention, Side, England, Women	R
74	Bulletin, Coronary, Therapeutics, Encyclopaedia, Angina, Technology, Assessment, Drug, European, Prevention, Heart, Action, Cochrane, Topic, Department	R
75	MZCP, Mediterranean, Organization, MZCPlow, States, Collaborating, Centres, National, Information, Power, Point, Greek, English, Region, Middle	I
IQ13	**Clinical, Stakeholder, National**	
76	Research, Stem, Centre, Biobank, Medical, Council, NIMR, Task, Force, Funding, Professor, Councils, Ethics, Governance, Member	R
77	Cancer, Trust, British, Association, Royal, College, Breast, Limited, Hospital, Health, Group, stakeholders, Centre, Research, Guideline	R
78	Disease, Department, BONE, Research, Dementia, University, Centre, Hospital, Medicine, Pagets, Ralston, ITALY, Belgium, Aberdeen, PRESS	R
IQ14	**Optic, Neuritis, Clinical**	
79	Ophthalmology, University, Research, Scase, Vision, Science, Foster, Bradford, Sciences, Professor, Hospital, Votruba, Heron, Journal, Optometry	R
80	Wednesday, University, Ophthalmology, Professor, Department, Cardiff, October, Optometry, November, Sciences, Vision, London, Hospital, Research, College	R

Table 9: Multiple Sclerosis (.uk) Cluster Labels (continued)

81	Journal, Ophthalmology, Research, University, Hull, Medical, Birch, Machin, British, Sciences, Multiple, Fitzke, Ophthalmol, Neurology	R
82	FAQs, Publications, Shop, Resources, Books, Internet, Information, Centres, Factsheet, Vision, Download, Publication, Reference, MS078, SEND	I
83	Impaired, Central, Fundal, Management, Refer	R
84	Journal, Influenza, Vaccine, Medical, Anonymous, Virology, Medicine, Health, Immunology, American, Review, General, Research, Pediatric, National	R

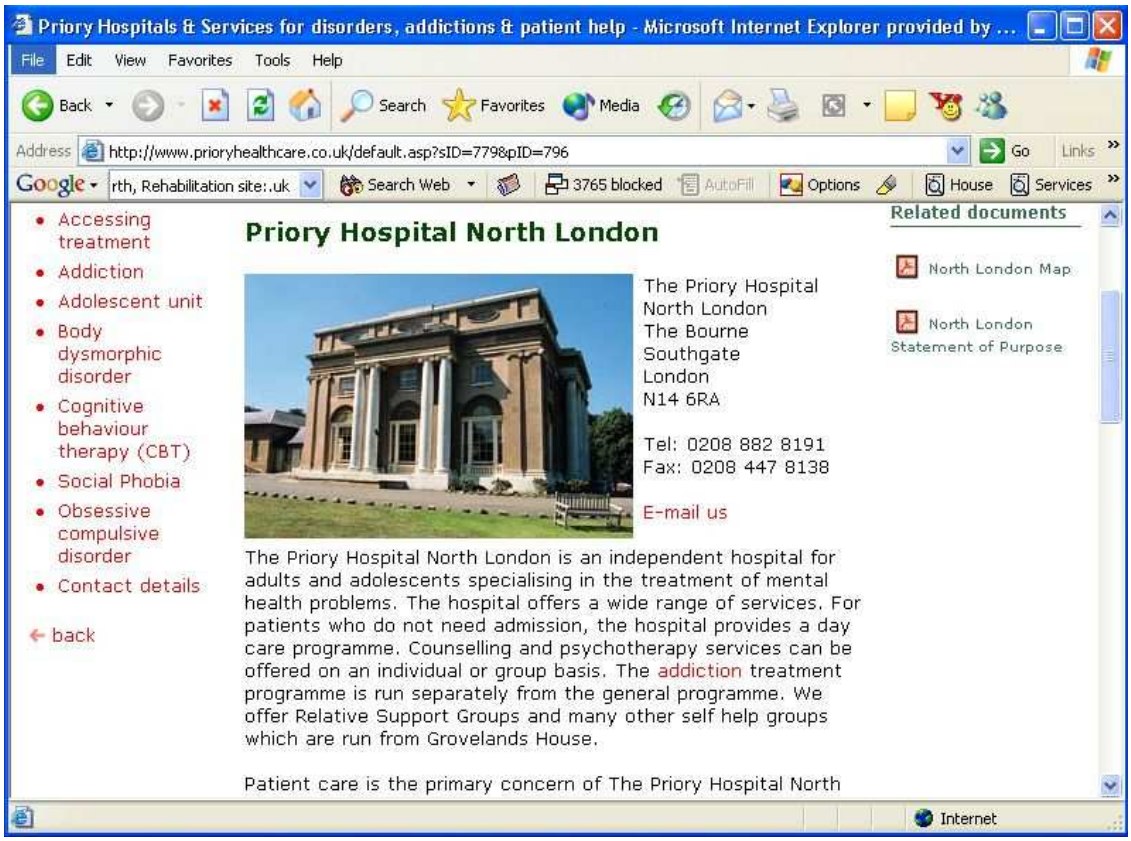

Figure 13: Web page for the query: *House Services Grange North Rehabilitation*

Table 9 identifies the novel clusters for this experiment with the labels for each cluster. Clearly most of the novel information retrieved concerns organizational entities such as hospitals, government sponsored educational programs for the disabled, charities, medicinal research among others. This information is indirectly related to *Multiple Sclerosis* in a manner that is sometimes not immediately obvious. For quite a few clusters, a search using the few nouns reveals interesting novel information. For instance, the top 5 nouns associated with cluster 63 (*House Services Grange North Rehabilitation*) do not seem to indicate an obvious relation with *Multiple Sclerosis*. However, a search on these nouns reveals highly relevant information about *Priory Hospitals*, which specialize in the treatment of mental disorders. The first web page retrieved using this search query is shown

Table 10: Expert Validation Statistics

	Number of Clusters	%age
Relevant	36	43
Partly Relevant	4	5
Irrelevant	39	45
Not–Thought–Of	3	5
Unknown	2	2
Total	84	100

in Figure 13. Taking another instance, the nouns in cluster 16 (Institute Grand Central Contact Foundation) seem non-cohesive. However, a search using these nouns inside the *.uk* domain reveals interesting information about *conductive education* in the first few pages.

The results obtained from this experiment are primarily directed towards an organizational perspective (as expected from governmental sites), such as information about the *Conductive Education* for people with disorders; drugs such as *Interferon*, *Copaxone* and *Bandolier*; charities for supporting patients suffering from disorders; *hospitals and support communities*, among others. Most of the clusters obtained possess a strong relevance to organizational work on the health and rehabilitation domains.

The results for this experiment have been validated with the help of a subject expert. Table 10

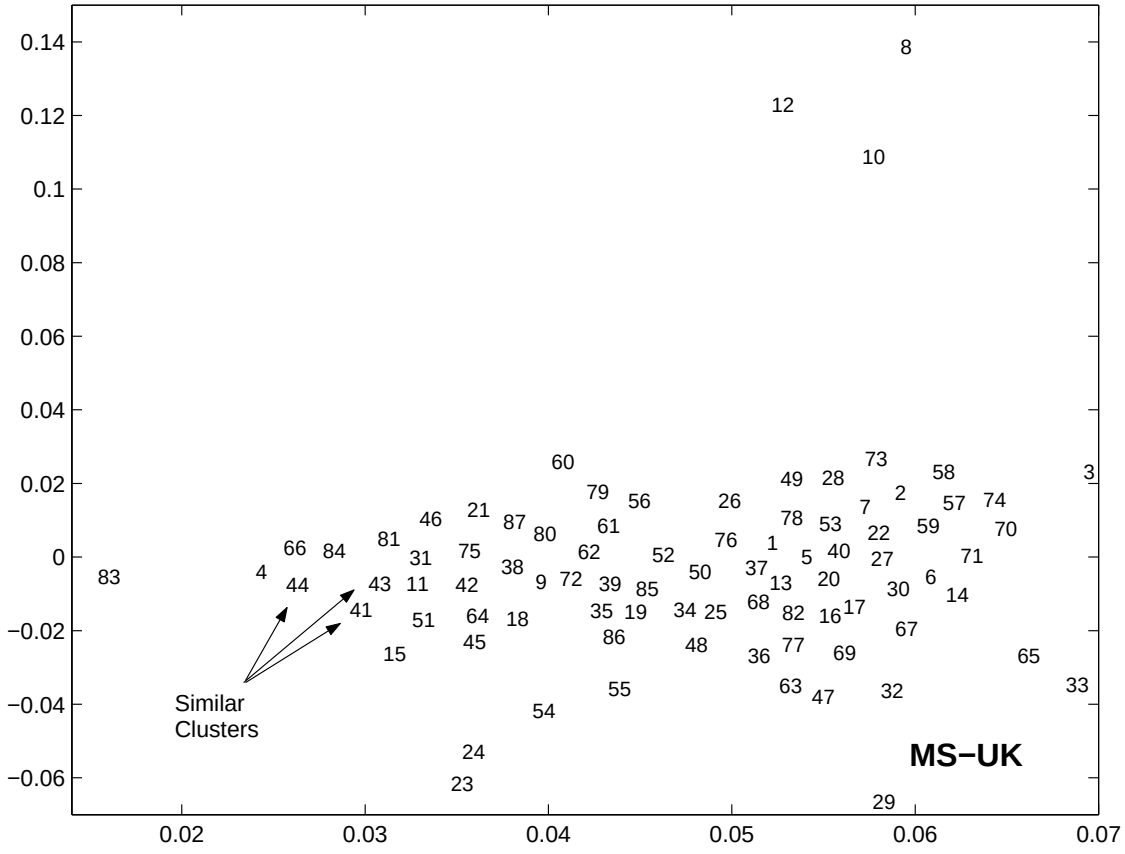

Figure 14: Novel Information Clusters

represents the statistics associated with the validation process. 43% of the total results were classified as relevant and 45% as irrelevant.

Figure 14 is the two-dimensional plot obtained after SVD on the novel clusters. Clearly the groups of novel clusters overlap significantly and no clear pattern is present.

4.7 Discussion

The use of 15-term descriptors captures the content of a cluster in a way that is useful for further search, but limited as a means of assessing cluster quality. Most of the clusters produced as the result of these experiments have considerable coherence. Indeed, these results show that there are well-formed clusters in the web that do not correspond to anything users would have created from an ontology perspective – and yet they exist.

It is harder to argue that the clusters produced are the appropriate novel information. It is certainly true that the clusters contain novel information; we have used other means to provide evidence that the novel information is, in some sense, good information. The results from human ratings suggest that human expectations and biases play a large role in determining the level of satisfaction with the results.

The results for the two 'multiple sclerosis' experiments show some of the interesting differences that arise when domains are restricted. Restricting the query to the .uk domain produces clusters whose content is about the organizational structure of medical care, clusters that do not appear when the query is applied to the entire web. This apparent emphasis on form as well as function may reflect the mindset of a government-run health system.

4.8 Conclusions

This paper proposed a systematic approach to novel information discovery using the ATHENS system. ATHENS attempts to move from known information to novel information in a directed way, by combining a firm starting point, contextualized query mechanisms, and powerful clustering of the results.

To the best of our knowledge, ATHENS represents the first systematic approach to solving the problem of novel information discovery in the Web. ATHENS does not require any background knowledge, creating its own by retrieving known information about the initial terms. The system can easily be adapted to other information resources for which the required underlying technology can be created.

The system was tested extensively, using different structural elements of the Web such as organizational domains, countrywide domains or the complete web. The quality of the results was assessed by determining their coherence and novelty, and by using subject experts to evaluate them.

The most promising benefit of the system is that it consistently retrieved novel information that possessed a high degree of interestingness for users. The novel information retrieved encompassed a wide range of topics that shared a high relevance to the query topic. The results also demonstrated a strong underlying context that related the novel information across a wide array of topics.

References

[1] M.R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, 1973.

[2] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[3] British National Corpus (BNC), 2004. www.natcorp.ox.ac.uk.

[4] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of 21st annual international ACM SIGIR conference on Research and Development in Information*, pages 335–336, 1998.

[5] F.R.K. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997.

[6] S. Deerwester, S.T. Dumais, G.W. Furnais, T.K. Landauer, and R.Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[8] M. Granitzer, W. Keinreich, V. Sabol, and G. Dosinger. WebRat: Supporting Agile Knowledge Retrieval through Dynamic, Incremental Clustering and Automatic Labelling of Web Search Results. In *Twelth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'03)*, 1080–1383, 2003.

[9] J2SE: Java 2 Platform, Standard Edition, 2004. **java.sun.com/j2se/1.4.2/download.html**.

[10] H. Liu. MontyTagger v1.2, 2003. web.media.mit.edu/hugo/montytagger.

[11] JAMA: A Java Matrix Package, 2004. **math.nist.gov/javanumerics/jama**.

[12] Penn Treebank Project, 2004. **www.cis.upenn.edu/ treebank/home.html**.

[13] D.B. Skillicorn and N. Vats. Novel information discovery for intelligence and counterterrorism. Technical Report 2004-488, Queen's University School of Computing Technical Report, September 2004.

[14] N. Vats and D.B. Skillicorn. Information discovery within organizations using the Athens system. In *Proceedings of 14th Annual IBM Centers for Advanced Studies Conference (CASCON 2004)*, October 2004.

[15] Google WebAPI. 2004. **www.google.com/apis**.

[16] G.K. Zipf. *Human Behaviour and the principle of least effort: An Introduction to Human Ecology*. New York. Hafner Reprint (1st Edition: Cambridge, MA: Addison-Wesley,1949), 1972.