PEx-WEB: Content-based visualization of web search results

Fernando V. Paulovich[1,2], Roberto Pinho[1], Charl P. Botha[2],
Anton Heijs[3], and Rosane Minghim[1]
[1]Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, São Carlos/SP, Brazil
[2]Faculty of Electrical Engineering, Mathematics and Computer Science,
Delft University of Technology, Delft, The Netherlands
[3]Treparel Information Solutions B.V., Delft, The Netherlands
paulovic@icmc.usp.br, rpinho@icmc.usp.br, c.p.botha@tudelft.nl,
anton.heijs@treparel.com, rminghim@icmc.usp.br

Abstract

The efficacy of search engines has expanded the uses for the information available on the Web. An increasing number of applications make use of the WWW as a primary source of information. The usefulness of such applications is, however, impaired by the current styles of display of the web search results. This paper presents a system that adapts two techniques to map and explore web results visually in order to find relevant patterns and relationships amongst the resulting documents. The first technique creates a visual map of the search results using a content-based multidimensional projection. The second techniques is capable of identifying, labeling and displaying topics within sub-groups of documents on the map. The system (The Projection explorer for the WWW, or PEx-Web) implements these techniques and various additional tools as means to make better use of web search results for exploratory applications.

Keywords—**Visualization and Knowledge Discovery, Visualization of Textual information, Web data Visualization, Interactive information retrieval, Text data mining**

1 Introduction

In most web search engines, the style of presentation does not motivate proper exploration of results related to a query. Lists can help locate better ranked results, but do not support extracting content relationship between the resulting set of documents. In applications that must make sense of more than just a few top ranked links (ex. patent search and exploration, analysis of news, forensic textual material and survey of scientific papers, etc.) better tools for analysis are required.

Visual mining (the coupling of mining and visualization algorithms) of web search results can help using search engines for much more than locating a particular link or document of interest. Various visualization approaches already exist to allow exploration of web pages and their content. Most of them are concerned with relating pages via common links or mutual citation. Although the existing approaches help to organize a search visually, there is currently little contribution to finding relationships and patterns that relate textual content.

Based on similarity relationships amongst textual documents, multi-dimensional projections can support exploration and understanding of the set resulting from a web search, particularly when coupled with an algorithm for topic extraction, such as the one implemented in PEx-Web. A projection technique maps the textual result from a web search (currently, it implements general search, patent search and RSS news feed retrieval) onto a 2D display that can be complemented with visual attributes. With the use of this tool, it is possible to explore groups of highly correlated textual content. Interaction tools, mining functionalities (eg. clustering and classification), and labeling from automatically extracted topics are available. These tools combined allow the user to achieve a broader and a detailed view of search results. Different queries can also be compared using additional tools implemented. It is our belief that the proposed approach represents a door to expanding the uses of web search results. In this paper we present the system, detail its framework, illustrate and evaluate its use.

The next section discusses work related to creating meaningful representations of search results and how they relate to the techniques and functionality of PEx-Web. Section 3 describes the functionality of the system. Section 4 describes experimental results and Section 5 shows the conclusions from a preliminary user evaluation. Additional Conclusions are drawn in Section 6.

1550-6037/08 $25.00 © 2008 IEEE
DOI 10.1109/IV.2008.94

2 Related Work

Today, most Web search engines return the results of a query to the user in the form of a ranked list of documents. Such representation has many limitations considering the number of applications that nowadays make use of the web as one of their primary sources of information. Other forms of display are necessary to reveal relationships not as linear as those offered by ranks. For some time now, various approaches have been proposed for organizing the search results in order to improve the process of finding useful documents [16].

Some of these approaches are based on the idea of creating visual representations of the retrieved documents [12], been suggested in [1] that these representations have significant benefits over a text-based interface. Currently, most systems that implement visualizations do so by creating a graph where nodes are pages and edges are the hyperlinks amongst them (see [9, 17, 5]). These are useful to place contents by their origin, an important type of relationship for various applications. This type of visual display is, however, not as useful when the goal is to explore results by of textual content, that is, the user wishes to associate pages by their similarity. For content-based tasks, an alternative form of visualization would be to connect place pages with high degree of common content. ReVEL [19] is a system where this approach is implemented. The main problem of ReVEL is that it ignores documents that are not HTML, thus a large amount of important documents formats are discarded.

Another way to help users comprehend search results is to split them into clusters and assign a label to each cluster. Currently there are two different approaches to perform this task [16]: the document-based approach and the label-based approach. In the former, clusters are created and representative term(s) or sentence(s) are extracted as labels from each cluster (ex: [4, 7]). In the latter, informative terms (words or phrases) are extracted from the search results as labels using a statistical analysis, and a cluster is defined by the documents that include a certain term (ex: [16, 20, 6]).

This clustering approach can effectively help the user to accelerate the task of browsing through the results. However, pre-clustering can have the adverse effect of loosing information on degree of similarity between elements within a cluster or between elements in different clusters.

Multidimensional projection techniques, employed by PEx-web, are capable of creating a visual representation of retrieved documents, mapping them as points on a plane where documents with highly correlated content are placed in the same neighborhood. While interpretation of the display is done by locating groups, there are no particular discrimination of their boundaries. Therefore, unlike the clustering approach, the intra- and inter-clusters relationships can be visually identified. By using text as source of information, this visualization is bound to represent a stronger semantic meaning than those using hyperlinks. The problem of document formats is handled in PEx-Web by taking as basis for processing the snippets (short descriptions of documents returned by search engines), thus being independent of document format whilst keeping the original document's URL.

The next section details the PEx-Web, presenting the features which are available to explore web search results.

3 PEx-Web: visually mining web results

The *Projection Explorer Web (PEx-Web)*[1] is an adaptation of a multi-dimensional visualization tool [11] to work with documents retrieved from Web. PEx-Web aims at supporting interpretation of collections of web results to avoid excessive visits to unwanted web pages. It does so by creating visual representations of retrieved documents and other supporting tools so that they can be explored by their correlation in content and other extracted relationships.

On the visual representations, called *document maps*, each document is represented as a point on a plane. Proximity amongst points indicates similarity of content. This representation is highly interactive and users can explore it to discover, for instance, relevant material to read, the subjects handled by groups of documents, and various other features.

Additional mining is accomplished by accessing the *KMX platform*, enabling data mining algorithms to be executed and their results to be shown as visual attributes on the document map. KMX is an integrated hardware and software solution to visual data mining developed by *Treparel Information Solutions B.V.*[2]. The system is a *solution in a box* where the client applications are as lightweight as possible regarding the computational power required; all of the complex tasks should be done on a remote server.

This first prototype of PEx-Web provides three ways to retrieve documents. One way is using the *Yahoo Web Search API* (`http://developer.yahoo.com/search/`). This API returns snippets for a query containing the same information that is returned on a search on the Yahoo web site. The second way to retrieve web documents is to provide a link to an RSS source and download the information associated with such link (for instance, flash news). It is possible to provide several links at the same time, and use the information of different sources to build up a single visualization. An RSS source provides short descriptions of a larger source of information, so it is also very fast to be retrieved. The third way is to select

[1] PEx-Web is implemented on Java and is available at `http://infoserver.lcad.icmc.usp.br/infovis2/PExWeb`.

[2] `http://www.treparel.nl`

links to RSS patent sources from the *Free Patents Online* site (`http://www.freepatentsonline.com/`).

The main window of PEx-Web is presented in Figure 1.

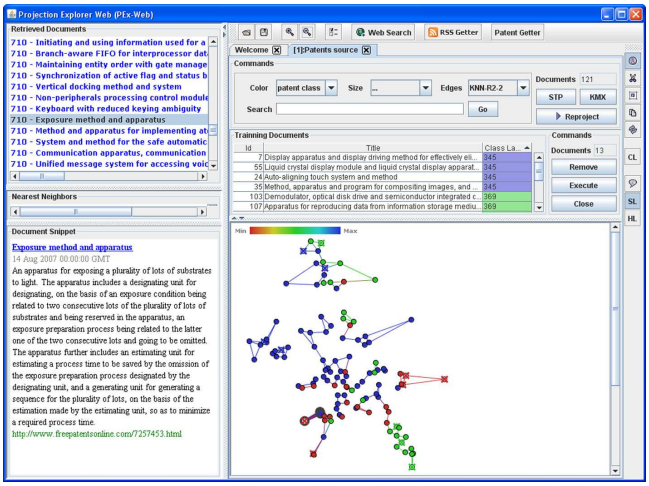


Figure 1: Main window of PEx-Web

3.1 Creating and exploring a document map

The process PEx-Web employs to create a document map is one used before ([11]) for text mapping based on content, and is summarized as follows:

- the retrieved documents are converted into a multidimensional vector representation (see Appendix A);
- a cosine-based distance metric [14] is used to calculate the dissimilarity between the documents as the distance amongst the vectors;
- a distance-based multi-dimensional projection technique, calles LSP, is applied to create the visual representation (see Appendix B).

Once a visual representation is created, it can be interactively explored.

If a user rolls the mouse over a point on the map, a label identifying its corresponding document is shown. A single mouse click over a point/document gives the user a list of its nearest neighbors. Neighbors on the projection are defined as the closest point or points according to Euclidean distance in 2D or according to thir similarity in the original domain. Selecting an area on the map opens all the pages of documents contained in that area.

Maps that contain many documents may result in a very cluttered graphical representation, thus such maps can be zoomed in or out; parts of the map can be selected and deleted.

Various queries can be submitted concurrently, and if two or more maps are open at the same time, it is possible to select an area of one map and verify if the documents contained in that area occur in the other maps. Thus, it is possible to analyze if queries with different search keywords return the same documents. Also, it is possible to join different maps into a single map created with all the distinct documents belonging to them.

One important tool to support exploration in this environment is the automatic extraction of topics dealt with by particular groups of documents. In PEx-web, the algorithm to extract topics is based on co-occurrence of terms. These terms are used as labels (see [3]). The projection can be submitted to clustering or classification and topic labels may be assigned automatically to the resulting groups.

Additional information can be mapped to the color and size of the points on the map using the frequency of a word or group of words in the documents. Also, for documents retrieved using the Yahoo API, the points can be colored or re-sized according to a rank of relevance defined by the engine; for the patent, points can be colored or re-sized according to the patent class given by the web site. The same visual attribute adjustments can be done to reflect classification or clustering of the text set.

In order to perform classification and clustering of the web search, we have used the KMX platform. For clustering, a hierarchical k-means algorithm is used and classification is done via the *Support Vector Machine (SVM)* technique (see Appendix C). For the clustering task, the user needs to define the number of clusters. For classification, the support vectors required by the SVM algorithm are chosen by clicking on points on the map. Good classification results have been achieved when choosing as support vectors points in groups that are well 'resolved', that is, groups that seem to contain documents relating to a common subject.

4 Results

In this section we present examples of maps created and explored using PEx-Web. In the next section we present results from a user evaluation conducted in order to verify user's ability to interact with the visual model and tools.

In the first example, a document map is created through a query on "visualization", a keyword which presents different meanings depending on context (Figure 2). On this map three different labels were created, showing different contexts where the word "visualization" can be employed: creative visualization, scientific visualization, and information visualization. It shows that PEx-Web can create graphical representations that separates groups of documents by content well, even when the keyword searched has a broad meaning or varying semantics.

In Figure 2, the documents are colored based on the frequency of the query "information AND visualization". It is possible to analyze the selected documents in further detail

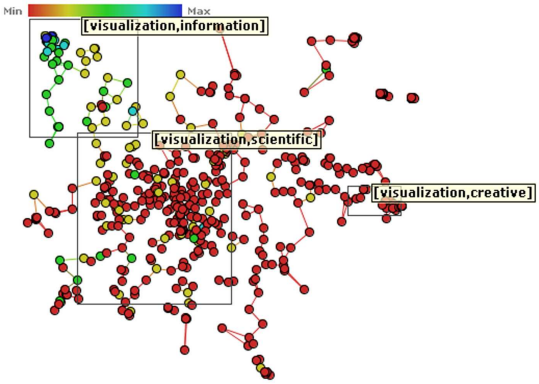


Figure 2: Example of a map using the keyword "visualization".

by removing the documents away from the area involved in the selection documents. This new map is presented in Figure 3. Observe other, not colored documents are employed, since their proximity is an indication of content similarity. Here, labels identify sub-topics related to information visualization.

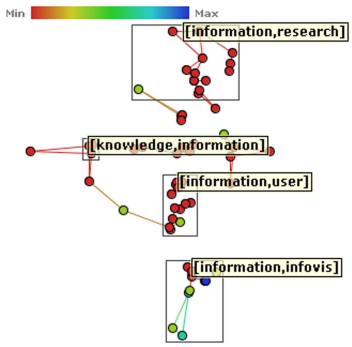


Figure 3: A new map generated using only a small selection of documents the the map on Figure 2.

Another interesting application of PEx-Web is the creation of document maps based on RSS sources. Figure 4 shows a map formed using the RSS sources provided by Associated Press (`http://www.ap.org/`), BBC (`http://www.bbc.com`), CCN (`http://www.cnn.com/`), and Reuters (`http://today.reuters.com/`). These news were collected on September 29th 2007, and the labels show some subjects of events occurred that day: liberation of Korean hostages in Afghanistan, persisting problems originated by hurricane Katrina, and so on. With maps of this type of data it is possible to locate subjects and events and also compare coverage of certain topics by different news agencies.

PEx-Web accesses the KMX platform for data mining

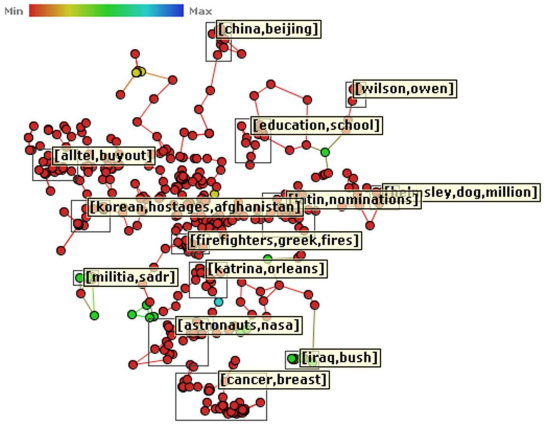


Figure 4: A document map with the Associated Press, BBC, CNN, and Reuters flash news.

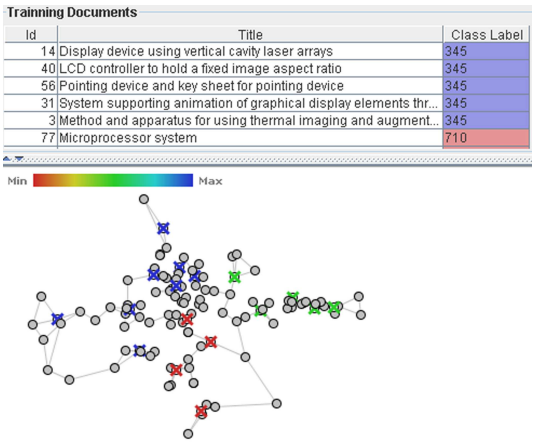
tasks. One of them is to perform SVM classification. Using PEx-Web the user chooses (by clicking) a subset of training documents and label their classes. The remaining documents are then classified using this subset as example. Figure 5(a) shows an example of training set selection using PEx-Web. The training documents are indicated by an 'X' over the corresponding point, and their color indicate the class assigned by the user.

The classification results are used to color the points on the map. In Figure 5(b), the documents are colored according to the classification given by the *free patents online* web site. The blue points indicate "graphic processing" patents, green points are "printers and printing" patents, and the red points are "I/O systems" patents. Figure 5(c) presents the final classification generated by the KMX platform with 5% documents in the training set. This classification matches almost completely the site's classes for patents. In PEX-web, for each identified class (or cluster) a topic is automatically extracted (the yellow boxes). In this figure, the "graphic processing" patents are identified as *(image,display)*, the "printers and printing" patents are identified as *(ink,jet)*, and the "I/O systems" patents are identified as *(data,bus)*, very consistent with the patents subjects.

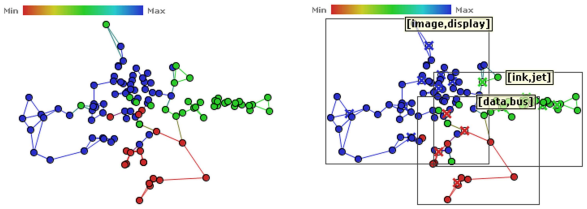
5 User Evaluation

We have performed a preliminary user evaluation procedure to help assess: (i) if users with little training could successfully use the system, (ii) if relevant documents are being properly presented to them, and (iii) if the tool could allow users to infer topics for a given region of the map.

Two tasks were devised: (i) a search task, in which users were asked to tag documents that were relevant to a given topic, and (ii) a topic evaluation task, in which users had to explore three colored regions and, for each one of these



(a) Selecting the training documents.



(b) *Free patents online* classification.

(c) KMX's SVM classification.

Figure 5: Training documents selection and examples of classifications provided by the *free patents online* web site and the KMX platform.

regions, select, from a list of topics, the one that the user considered most related to the documents found on that region (Figure 6).

To perform the experiment, users were asked to follow instructions from pages reachable from the PEx-Web page, where a short (5 minutes) instruction video and specific versions of PEx-Web were available. Every user should: (i) read general instructions, (ii) answer quick profile questions, (iii) watch the video, (iv) perform the search task, and (v) perform the topic evaluation task.

A few standard options from PEx-Web were disabled as users explored the same preloaded maps. Disabled options were kept visible in order to keep the look and feel as close as possible to the standard version.

Both tasks used a selection of news articles extracted from the Reuters Corpus [8]. A base corpus was built by selecting documents judged as relevant for at least one topic of the TREC 2002 filtering task training set[3]. For the search task, documents from the test set were added until the number of relevant documents for the specific topic of "Effects of global warming" reached a number between

[3] `http://trec.nist.gov/data/t2002_filtering.html`

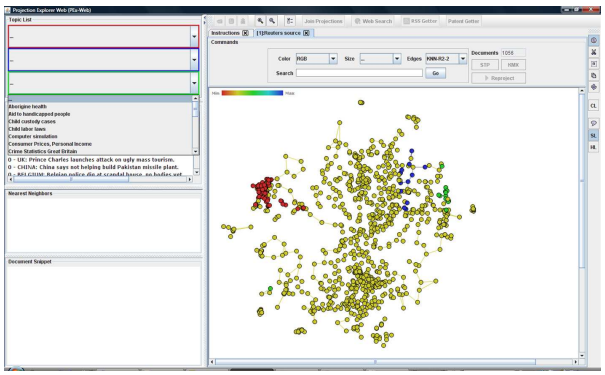


Figure 6: A document map for the topic task with three colored regions and partial list of topic alternatives shown.

10 and 20. This specific topic was chosen by a researcher that is not a member of our group with only one guideline: choose a topic that interests you from the list of topics retrieved from the TREC task. The same procedure was performed to build the topic task corpus, but this time with three different topics: "Rescue of kidnapped children", "Improving aircraft safety" and "Progress in treatment of schizophrenia". Documents considered relevant on the TREC task for these topics were colored red, green and blue respectively on the map. The final corpora have 1045 documents for the search task, and 1056 for the topic task. Other topics from the TREC task were chosen for being displayed as alternatives for the answers on the topic task. This means that the alternatives were also represented on the map.

Maps were built using the tool's default options and the first produced maps were adopted.

For the search task, users had high recall (avg: 52.0% std. dev.: 40.1%) and precision (avg: 83.8% std. dev.: 37.3%). Almost half (3 out of 7 users) had a recall of 92.9%, thus confirming that users with little training can successfully use the system. Every relevant document was tagged as relevant by at least one user, which means that the system did not hide any relevant document or made unfeasible for a relevant document to be reached. For the topic task, users had 16 right answers out of 18 (3 answers $\times$ 6 users). The only two errors were related to the topic "Rescue of kidnapped children". In both cases, the users had selected this answer but latter changed their minds. This last result shows that the system allows users to infer topics for a given region.

6 Conclusions and Further Work

We have presented PEx-Web, the Projection Explorer for the Web, a highly interactive tool to support exploration of textual information retrieved from the WWW. PEx-Web employs recent developments in multidimensional projec-

tion technique and topic extraction to create a visual representation, a document map, where documents are represented as points on a plane, similarity of documents is identified as proximity on the plane, and labeling is done by automatically extracting topics. A reasonably large number of documents can be analyzed in one document map.

In Pex-Web, the maps can be created from queries or RSS feeds (particularly, news and patents are pre-set). The projection based visualizations tend to group together results with high content correlation and to separate groups related to distinct subjects. PEx-Web includes a technique to extract topics of groups of text automatically.

We have tested PEx-Web in various different scenarios, and in all cases it suggests the approach can decrease the time that the user spends to analyze the results of a web query in order to find useful documents insights, particularly for exploratory applications. Processing and display are fast and allow exploration in real time. We performed a more formal evaluation with non-specialized users to confirm that they can successfully understand and use the tool. Other, broader tests are currently being done.

Acknowledgments

This work is supported by FAPESP research financial agency, São Paulo, Brazil (proc. no. 04/07866-4 and 04/09888-5), and CAPES research financial agency, Brazil (proc. no. 2214-07-5). We wish to acknowledge the work of our undergraduate and research students as well as research colleagues in processing some data and discussing various issues of the work.

References

[1] A. Becks, C. Seeling, and R. Minkenberg. Benefits of document maps for text access in knowledge management: a comparative study. In *SAC '02: Proceedings of the 2002 ACM symposium on Applied computing*, pages 621–626, New York, NY, USA, 2002. ACM.

[2] N. Cristianini and J. S. Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[3] A. M. Cuadros, F. V. Paulovich, R. Minghim, and G. P. Telles. Point placement by phylogenetic trees and its application for visual analysis of document collections. In *IEEE Symposium on Visual Analytics Science and Technology 2007 (VAST 2007)*, pages 99–106, 2007.

[4] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 76–84, New York, NY, USA, 1996. ACM Press.

[5] KartOO. `http://www.kartoo.com`.

[6] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 658–665, New York, NY, USA, 2004. ACM Press.

[7] A. Leuski. Evaluating document clustering for interactive information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 33–40, New York, NY, USA, 2001. ACM.

[8] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[9] S. Mukherjea and Y. Hara. Visualizing world-wide web search engine results. In *IV '99: Proceedings of the 1999 International Conference on Information Visualisation*, pages 400–405, Washington, DC, USA, 1999. IEEE Computer Society.

[10] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, 2008.

[11] F. V. Paulovich, M. C. F. Oliveira, and R. Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *SIBGRAPI '07: Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, pages 27–36, Washington, DC, USA, 2007. IEEE Computer Society.

[12] R. M. Rohrer and E. Swing. Web-based information visualization. *IEEE Comput. Graph. Appl.*, 17(4):52–59, 1997.

[13] G. Salton. Developments in automatic text retrieval. *Science*, 253:974–980, 1991.

[14] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.

[15] E. Tejada, R. Minghim, and L. G. Nonato. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, 2(4):218–231, 2003.

[16] H. Toda and R. Kataoka. A search result clustering method using informatively named entities. In *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 81–86, New York, NY, USA, 2005. ACM Press.

[17] H-C. Yang, M-C. Tzeng, and C-Z. Yang. A web interface for visualizing web search engine results. In *Proceedings of the 3rd International Conference on Parallel and Distributed Computing, Applications, and Techniques (PDCAT 2002)*, pages 283–288, 2002.

[18] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[19] C. M. Zaina and M. C. C. Baranauskas. Revealing relationships in search engine results. In *CLIHC '05: Proceedings of the 2005 Latin American conference on Human-computer interaction*, pages 120–127, New York, NY, USA, 2005. ACM Press.

[20] H-J. Zeng, Q-C. He, Z. Chen, W-Y. Ma, and J. Ma. Learning to cluster web search results. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2004. ACM Press.

A Documents Processing

The process employed here to define the similarity amongst the documents initially converts the set of documents into a vector representation. In this vector representation, each document is converted into a vector with coordinates based on the frequency of terms. After that, these vectors are united to form a matrix of documents x terms in which each term is weighted according to the *term frequency inverse document frequency (tfidf)* [13].

In order to reduce the typical high dimensionality of such data set, we have employed a slightly modified version of the *document frequency threshold* approach [18]. This approach is based on the number of documents in which a term occurs. Here, if a term appears in more than 50% and less than 5% of the documents it is discarded. In this way, terms that are not useful to distinguish documents are discarded.

B Multidimensional Projection

A multidimensional projection technique is a injective function $f : \mathbb{R}^m \rightarrow \mathbb{R}^p$ which maps data from an m-dimensional space into a p-dimensional space with $p = \{1, 2, 3\}$ and $p < m$, whilst retaining, on the projected space, some information about distance relationships among the data items in the m-dimensional space. Formally, let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of m-dimensional data, with $\delta : \mathbb{R}^m, \mathbb{R}^m \rightarrow \mathbb{R}$ a dissimilarity measure between two m-dimensional data instances, and $d : \mathbb{R}^p, \mathbb{R}^p \rightarrow \mathbb{R}$ a distance (normally Euclidean) between two points of the projected space. A multidimensional projection seeks to make $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$ as close to zero as possible, $\forall x_i, x_j \in X$ [15].

PEx-Web employs the *Least-Square Projection Technique (LSP)* [10] to create the document maps. LSP is a non-linear projection technique that adapts an approach for mesh-recovering and mesh-editing in order to deal with high dimensional spaces. In this technique, a subset of m-dimensional points are projected onto the plane, and the remaining points are projected using an interpolation strategy that considers only the neighborhood from the m-dimensional points.

C Support Vector Machine

Support Vector Machine (SVM) [2] is a supervised classification method which aims at obtaining a classifier that minimizes the training set error and the confidence interval, which corresponds to the generalization or test set error.

In SVM, the idea is to fix the training set error associated with an architecture and then to use a method to minimize the generalization error. The primary advantage of SVM as adaptive models for binary classification is that they provide a classifier which implies low expected probability of generalization errors.

For classification, SVMs operate by finding a hypersurface in the space of possible inputs. This hypersurface will attempt to split the positive examples from the negative examples. The split will be chosen to have the largest distance from the hypersurface to the nearest of the positive and negative examples. Intuitively, this makes the classification correct for testing data that is near, but not identical to the training data. In machine learning, this distance is referred to as "margin", and the classification surface we are looking for is therefore the "maximum margin hyperplane".

Currently, the SVM method is considered the only approach that is computational efficient and for which there is a well-defined theory which describes the mechanisms with respect to its accuracy and robustness.