RC22174 (W0109-020) September 17, 2001
Computer Science

IBM Research Report

Visualizing the Patent Document Collection as a Graph of Inventors and their Inventions

Douglas N. Gordin, Robert G. Farrell
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g. , payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

Visualizing the Patent Document Collection as a Graph of Inventors and their Inventions

Douglas N. Gordin Robert G. Farrell
IBM T.J. Watson Research Center
{dgordin, robfarr}@us.ibm.com

Abstract

Large document collections pose serious cognitive challenges for users attempting to find documents related to their interests and discover global relationships and groupings. We describe here the Raisin system that provides a better way to view, group, and analyze the results of searching a document collection, with a particular focus on patents. In particular, Raisin provides facilities to: (A) Select a sample of the document base including keyword search and crawling; (B) Create and dynamically lay out two mode graphs showing authors and their documents; (C) Generate short descriptive labels for the documents; (D) Collect and display focused indices; and (E) Simplify graphs by aggregating vertices (or folding them). These techniques support the navigation of document collections by visualizing and analyzing the implicit social networks found in authors and their documents.

1. Introduction

Large document collections pose a serious cognitive challenge for users attempting to find documents related to their interests and discover global relationships and groupings. Typical techniques for helping users include indices of document meta-data such as authors, keywords, and titles, as well as, classification taxonomies and reverse citation indices. However, the interfaces that use these structured information sources often merely present search results as numbered lists, thus obscuring relationships amongst the selected items. The document collection of U.S. patents (http://www.uspto.gov/) has over 28 million entries including author, patent assignment, title, and classification. The search engine at the US Patent and Trademark Office implements a sophisticated search on these fields, but presents the results as a simple list.

2. Visualizing Document Collections

We describe here the *Raisin* system that provides a better way to view, group, and analyze the results of searching a document collection, with a particular focus on patents. Prior work that helped motivate this approach includes [2] and [4]. The primary techniques that *Raisin* provides to enable document collection exploration are:

(A) *Selecting a subset of the Patent document collection by specifying a keyword and/or traversal (crawl) of index fields*. This sample can be specified using a set of keywords and/or instructions about crawling. For example, keyword phrases such as "Information Visualization and Graphs" can be used to select a set of patents (see Figure 1). Crawling extends this set by including the inventors' other patents or including referenced patents. The set can be expanded indefinitely in this way (e.g., including the patents referenced by a patent reference and so on).

(B) *Graph of inventors and inventions that shows the results of a patent document collection search and/or crawl*. The set of selected patents are shown by a graph that includes vertices for inventors and their inventions with edges indicating authorship. The graph layout is computed so that the distance between vertices (i.e., the Euclidean distance) is proportional to the graph distance (i.e., the shortest number of hops to go from one vertex to another).

(C) *Automatic derivation of vertex labels*. Patent vertices are labeled with the two most significant words or word pairs that occur in their text. These words or phrases are picked from a patent's title by contrasting its text with the patent corpus.

(D) *Focused indices of selected inventors and inventions*. Inventors are indexed by their affiliation, as indicated by the ownership of the patent (i.e., assignee). Inventions are indexed by the classifications. These indices are provided alongside the patent graph.

(E) *Graph folding by index*. The graph is simplified by replacing a set of vertices (representing inventors or inventions) by a new vertex that represents them in aggregate (see Figure 2). Only the new aggregate node continues to appear, the folded ones are hidden. The new node inherits the edges of the ones it replaces. The user chooses a portion (or all) of the index to aggregate using a twist-down control.

3. Example Graph Visualizations

This section demonstrates the *Raisin* system using the keyword query: "information visualization and graph." (Figure 1). There are three types of vertices: The search expression (shaded pink), the patents (shaded green), and the inventors (shaded yellow or blue). The index for

inventors appears in a tree control on the left. The currently selected organization is "Silicon Graphics." Therefore the Silicon Graphics inventors are shaded blue, rather than the yellow used for non-selected inventors. The patents are labeled by two or three word phrases that are selected by a basic text-mining algorithm. A term is ranked using the ratio of its frequency in the patent and its frequency in the patent corpus (i.e., term frequency multiplied by inverse document frequency [3]). The full title can be seen in a bottom pane on mouse-over and the complete patent text accessed by a double click. For example, the "directed graph" label refers to a patent titled, ""Graphical user interface for displaying and navigating in a directed graph structure". The layout of the graph is automatically computed using a version of Cohen's multi-dimensional scaling approach [1]. This approach computes the stress on a vertex by calculating the difference between its graph distance to the other vertices and its Euclidean distance to the other vertices. Each vertex is then moved by a fraction of this amount, thereby, over repeated iterations, seeking to minimize the total stress of all the vertices.

In Figure 2 you see a folding operation that results in the individual inventors being represented by the companies to which they assigned their patents. This is a powerful visualization because it allows the viewer to begin to categorize the various ways that the topic being investigated "information visualization" is being pursued at various companies. In particular, a viewer might conclude that Silicon Graphics is a major player by virtue of its high degree (outgoing edges); Mercury Interactive is focusing on Web applications; and Lucent has developed 3D network views.

4. Conclusions and Future Work

Communities of practice organize their work through document collections. Graphing these collections as two-mode graphs of authors and their documents provides a powerful tool for finding connections and aiding navigation. The Raisin system produces graphs of this sort for the patent document collection. Further, it enhances their utility by inferring text labels, providing indices, and performing graph folding. In future work, we will suggest graph foldings, provide ways to dynamically modify indices, provide indices based on graph structure and lexical content, use indices for hiding vertices (as opposed to aggregating), and provide graph analyses.

5. References

[1] J.D. Cohen. "Drawing graphs to convey proximity: an incremental arrangement method." *ACM Trans. Comput.-Hum. Interact. 4*, 3 (Sep. 1997), pp. 197 – 229.

[2] H. Kautz, B. Selman, M. Shah, "Referral Web: Combining Social Networks and Collaborative Filtering." *Communications of the ACM, Vol 40*, No 3. March 1997.

[3] Moens, M.F. *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers:Boston, MA. 2000.

[4] R. Xiong, M.A. Smith and S. Drucker. "Visualizations of Collaborative Information for End-Users", *Microsoft Technical Report No. MST-TR-98-52*, October 1998.

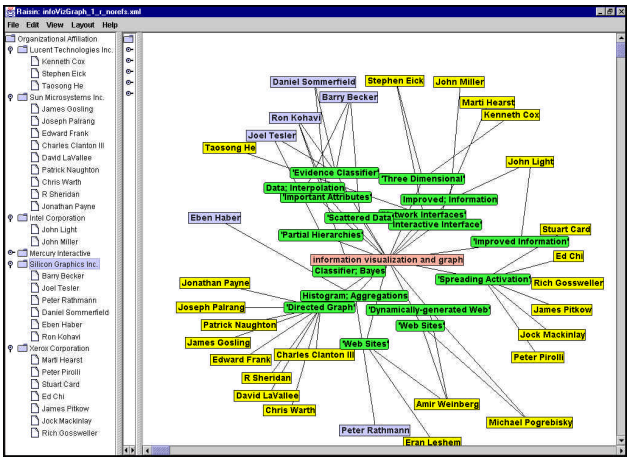

Figure 1: Graph of "information visualization and graph".

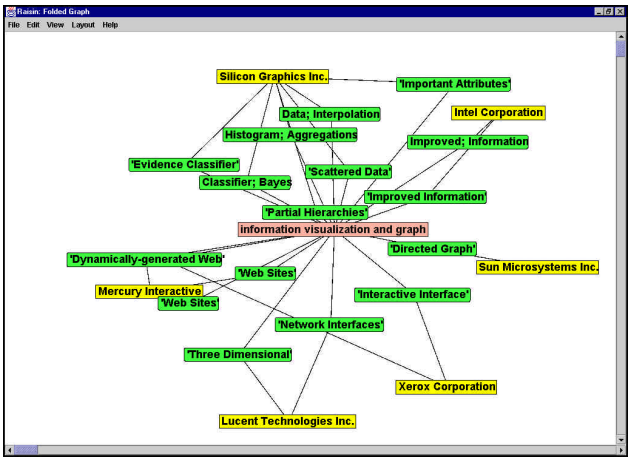

Figure 2: Folding by inventor's organizational affiliation.