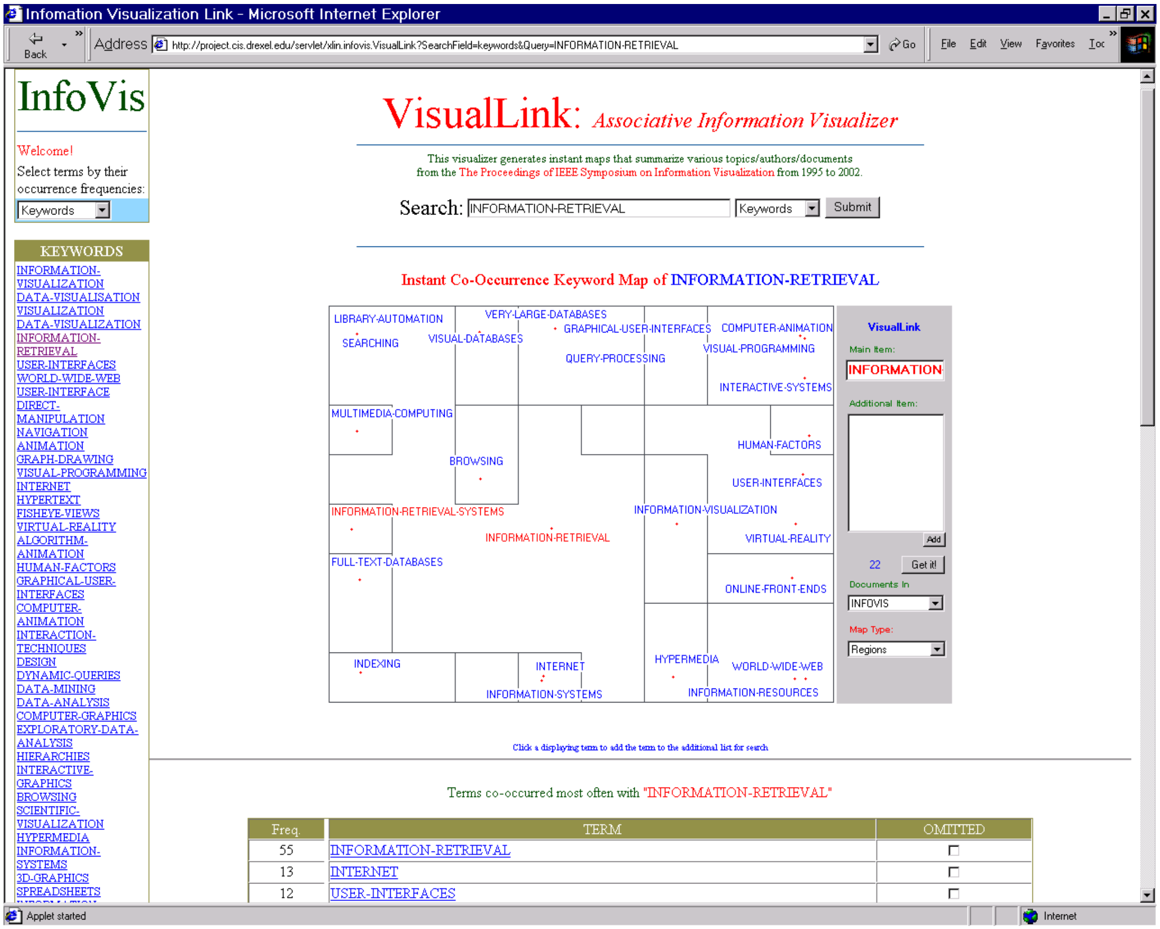
An Associative Information Visualizer

Howard D. White[1] Xia Lin[2] Jan Buzydlowski[3]
College of Information Science and Technology, Drexel University



Infomation Visualization Link - Microsoft Internet Explorer

Address http://project.cis.drexel.edu/servlet/xlin.infovis.VisualLink?SearchField=keywords&Query=INFORMATION-RETRIEVAL

InfoVis

Welcome!
Select terms by their occurrence frequencies: Keywords

VisualLink: *Associative Information Visualizer*

This visualizer generates instant maps that summarize various topics/authors/documents from the The Proceedings of IEEE Symposium on Information Visualization from 1995 to 2002.

Search: INFORMATION-RETRIEVAL Keywords Submit

Instant Co-Occurrence Keyword Map of INFORMATION-RETRIEVAL

VisualLink
Main Item: INFORMATION
Additional Item:
22
Documents In: INFOVIS
Map Type: Regions

Click a displaying term to add the term to the additional list for search

Terms co-occurred most often with "INFORMATION-RETRIEVAL"

Freq.	TERM	OMITTED
55	INFORMATION-RETRIEVAL	□
13	INTERNET	□
12	USER-INTERFACES	□

Applet started

Figure 1. A self-organizing map of keywords associated with the seed term “Information Retrieval”

1 Introduction

As a continuation of our previous work in the mapping of terms from the bibliographic records of scholarly and scientific literatures [1-2], we have created VisualLink, an associative information visualizer, and applied it to content from the InfoVis 2004 dataset. A sample page appears as Figure 1, to be discussed after we state a few of our design principles:

- Associated terms are those that frequently co-occur. The terms most worth mapping are those that rank highest in frequency of co-occurrence.
- Maps should reveal term associations in big full-text or bibliographic databases that reflect real-world literatures.
- Both Pathfinder networks (PFNETs) and Kohonen self-organizing maps (SOMs) are useful ways of displaying top-ranked term associations in two dimensions [3-4]. SOMs show the ranking of associations by the relative proximity of points standing for terms (the higher the ranks, the closer the points), while PFNETs explicitly link pairs of points for terms that most frequently co-occur.
- Users should be able to create maps with a single seed term or phrase, thereby minimizing the input needed to use the system.
- Maps should reveal interesting associations hidden in the database, such as the authors cocited with an author or the keywords co-occurring with a keyword.
- To maintain users’ momentum, maps should appear quickly after the seed term is entered.
- Cognitively speaking, SOMs and PFNETs with 25 to 50 term points are about the right size. They limit a global

[1] e-mail: whitehd@drexel.edu
[2] e-mail: xlin @drexel.edu
[3] e-mail: sg81qbdh@drexel.edu

domain to the semantic neighborhood of the seed. They portray a rich but not overpowering set of relationships.

- Term labels for 25 data points can be placed in 2-D space with little or no overlap, which permits the swift assessment of relationships.
- Users should be able to "cross-map" domains—that is, to find a seed term's associations not only with terms of its own kind but with terms of a different kind. For example, it should be possible to translate a seed author into associated keywords or a seed keyword into associated authors.
- Since the mapped terms index bibliographic or full-text databases, the maps should be capable of serving as live interfaces for the retrieval of items from the databases.

All of these design principles have been implemented in VisualLink, which displays data from the bibliographic records of 615 InfoVis *Proceedings* papers that appeared in 1995–2002. The records were parsed and manipulated with the Python programming language. The data extracted were primary authors, cited authors, cited documents, publication years, keywords (and later INSPEC keywords), and stemmed noun phrases from abstracts and titles. The output was placed in Noah, a specialized database that we created for real-time "two-by-two" processing of co-occurrences.

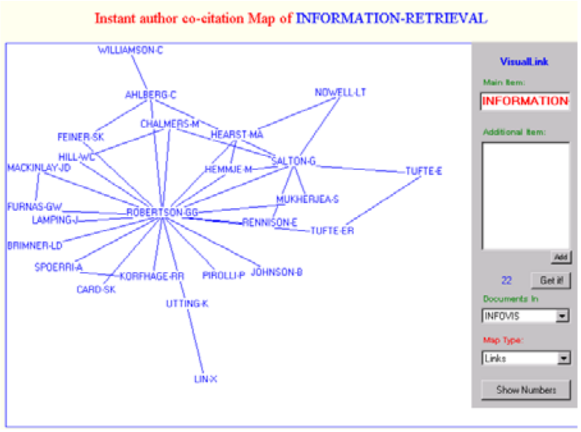
2 Discussion

In Figure 1, the left panel is topped by a pull-down menu for selecting seed-term type. It is here set for Keywords. Other choices are Authors, Cited Authors, References (i.e., cited documents), and "Overview," an option that produces an interactive subject map of the full *Proceedings*. The choice of Keywords has called forth a panel of suggestions, from which Information Retrieval has been picked. This action automatically places it as seed in the Search window at top center of the display. To the right of that window is another pull-down menu, this one for choosing the type of terms that the seed term will summon to be mapped. Again, Keywords has been picked, and the Submit button has generated the map, a Kohonen SOM that shows the seed term surrounded by the 24 INSPEC subject terms most closely associated with it. Other choices from the central menu will lead to "cross-mappings": the authors or the documents most heavily cocited in papers indexed with Information Retrieval, the stemmed terms that co-occur most heavily in their titles and abstracts, or a mixture of all of these. Moreover, users can create other new maps instantly with terms obtained in the present mapping.

Figure 1's gray panel at right has a menu for switching map types from SOMs ("Regions") to PFNETs ("Links") and back. It also includes the window for assembling terms from the map for retrieval of documents from the underlying database (here, for test purposes, the ACM Portal). Any mapped term can be placed in it by point-and-click. The term(s) will be automatically ANDed with the seed term in a search, which is executed with the Get It button. In blue nearby is a count of how many documents a particular search formulation will retrieve (e.g., 22). The small Add button allows the user to add to a search any non-mapped term the database may support.

Partially visible at the base of Figure 1 is a large panel that rank-orders the top 50 terms co-occurring with the seed term. With it, users can remap a particular domain after unwanted terms have been removed or require a seed term to appear in all combinations of terms when the counts for a map are obtained, which is a strong contextualizing feature.

Figure 2 shows one "cross-map." It is a PFNET of the 25 most cited author-names in the Information Retrieval set. The links indicate who has the *highest* (or tied highest) cocitation counts among all pairs of that set (the Show Numbers button puts the counts above the links). Obviously in IR as construed by InfoVis contributors, G.G. Robertson is a major influence. Other names will vary in familiarity to readers literate in the IR domain.



Instant author co-citation Map of INFORMATION-RETRIEVAL

Main Item: INFORMATION

Additional Item:

22

Documents In: INFOVIS

Map Type: Links

Figure 2. Pathfinder network of cocited authors in IR

The strength of both SOMs and PFNETs is that they heighten awareness of vocabularies that will be fruitful in retrievals and yield insights into term interrelationships. For example, the SOM algorithm automatically indicates that, in the InfoVis world, Information Retrieval implies work in graphical interface design, as shown by terms clustered in the right half of Figure 1. It implies the Internet, as shown by terms clustered along the bottom. And it implies traditional library-related indexing and searching in large textual and multimedia databases, as shown by terms clustered at left. Given this presentation, users can *recognize* terms they need without having to do lookups in thesauri or directories. Similarly, Figure 2 highlights authors who contribute ideas to IR in the InfoVis world. Some of them may be unknown to persons newly browsing this domain. Figure 2 invites exploration as to why certain pairs of authors—e.g., Robertson and Marti Hearst—are related.

The main weakness of the maps is the generality and ambiguity of the keywords, authors' names, or other vocabulary in them. Even when they are clustered or linked, it is not readily apparent what the terms in fact mean or what they will bring forth when used in retrievals. But this is a weakness of indexing in general; and we do not think it keeps VisualLink from frequently being informative.

References

[1] Howard D. White, Xia Lin, Jan W. Buzydlowski, Chaomei Chen. User-controlled mapping of significant literatures. *Proceedings of the National Academy of Sciences* 101 (suppl. 1): 5297-5302, April 6, 2004.

[2] Xia Lin, Howard D. White, Jan Buzydlowski. Real-time author co-citation mapping for online searching. *Information Processing and Management* 39: 689-706, 2003.

[3] Roger W. Schvaneveldt, ed. *Pathfinder associative networks: Studies in knowledge organization.* Ablex, Norwood, NJ, 1990.

[4] Teuvo Kohonen. *Self-organizing Maps.* 2d ed. Springer, New York, 1997.