

On Key Bespoke Tools to Support Electronic Academic Document Discovery

Fernando Loizides^{a,1}, George Buchanan^b and Keti Mavri^a

^a*Cyprus University of Technology*

^b*City University London*

Abstract. Publishing in academic journals and conferences has become faster, and easier with the ability to edit and submit documents electronically. With the increase of publications also come negative effects such as that of information overload and elevated discovery time of relevant resources. An information seeker often wades through several documents in order to find relevant publications having to either select known repositories for their search or utilizing generic search sources which network to several online repositories. Even with the advances in interactive systems, information seekers still carry out a mostly textual search from input to returned results. Several tools have been created by researchers in order to assist the seekers in their visual academic document triage activities but very few have been successfully implemented in actual discovery of electronic publications. With electronic publishing increasing dramatically, we recognize the paramount importance for these tools to be improved and integrated within environments to assist the seekers. In this work, we present an overview of key bespoke tools purpose built for achieving this document selection tasks. Using this work as a reference we hope to encourage structured and novel approaches to creating triage tools and improve the discovery process of electronic academic document publications.

Keywords. Document triage, tools, information seeking

1. Introduction

Electronic publications continue to increase exponentially with the advent of new publishing routes; namely, increasing conferences, journals and online document repositories. Document discovery is becoming more difficult, with the key focus being on the process of publication and retrieval techniques for single database recovery. What is currently under researched and under supported is the process of discovering these publications; a stage which renders the actual publication process void if the documents are never read. Academic repositories are now increasing the tools available to users in order to discover information within documents as well as documents themselves (See Figures 1 and 2).

The purpose of this work is to give the reader a directed primer on the types of tools that have been created over the years to support the document selection and discovery process. We do not aim to present an exhaustive literature review, but rather key findings which represent the under researched field, in order to equip developers, designers and stakeholders with the information needed to guide informed decisions on research and development; crudely speaking, a starting point for the masses. We bring

¹ Corresponding Author. E-mail: fernando.loizides@gmail.com.

together a body of work which is very focused and directed and present suggestions with a hope to encourage more work to be undertaken with the correct incentives in the field of document selection and interfaces to support information seekers.

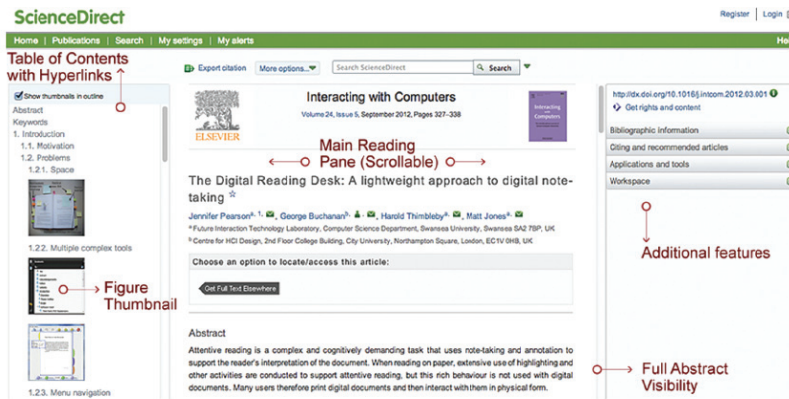


Figure 1. Science Direct website providing tools for faster navigation, also assist information seekers in their triage activities

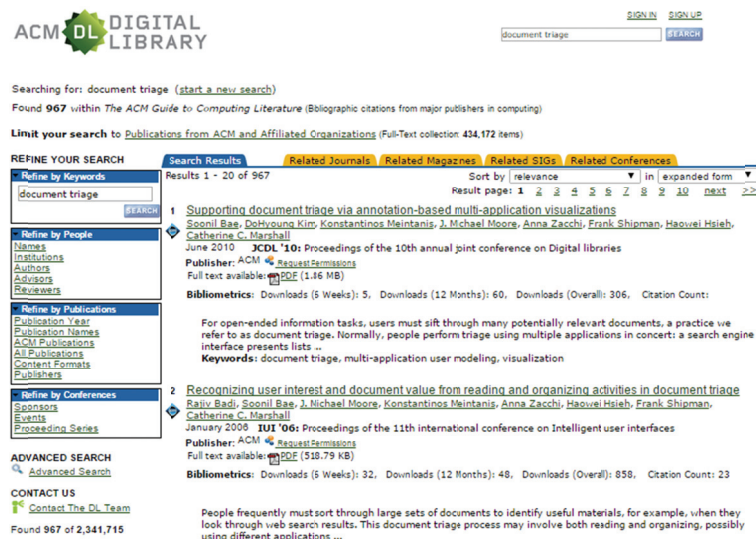


Figure 2. The ACM Digital Library website providing tools for document searching and discovery through facets and snippets

Document Selection is a process undergone by scholars, information professionals and information seekers daily to choose relevant documents on a topic. More recently, the term document triage has been adopted to describe more accurately the document selection process. Document triage is largely a human cognitive process and has not been thoroughly researched, hence this process is not yet fully understood. In order to understand the effect that document triage has on information seeking we focus on the

part of the information seeking process that document triage influences. Information seekers are reported on as making erroneous decisions on the relevance of documents during triage [1]. When influencing factors, such as document features, are altered the behavioural patterns of the users are also likely to change. There are three levels at which the triage process can take place [2], the surrogate stage, the within-document stage and the in depth reading stage.

Most of the work on document triage has been using manual searching, without specialized support from software. Research reports indicate how the libraries themselves lack “better support users’ overall information work in context” [3, 4]. Some work, albeit limited, has been carried out to investigate how supportive software can assist users in their information seeking activities. In this work, we take a closer look at key bespoke tools created by researchers and how they affect the document triage process. We begin with an overview of information visualization, which is a technique that is employed in the majority of the individual tools presented. We then present the individual tools themselves. We then discuss the tools as a whole, their potential and limitations. We conclude with future directions that can be taken related to the area of discovering academic publications with tools.

2. Tools and Concepts

2.1. Information Visualization

One of the most challenging problems performing document triage from a results is the sheer amount of documents available. An information seeker is often inundated by more documents that can be possibly looked at. One way that researchers attempt to solve this problem is by using information visualization within their proposed tools. We therefore give a small primer to the reader.

Information visualization has not been restricted to the visual cues alone, but has evolved to include the interactions with the information [5]. Visualizations have, thus far, mostly been effective in a more structured or hierarchical form [6, 7, 8]. Research into query tools, utilizing visualizations to a search a document corpus, has been conducted with positive results [9, 10]. Of course, visualizations are not without their challenges [11], but the results reported are mainly positive and outweigh these issues. Furthermore, advances in information retrieval algorithms (like the TREC conference [12, 13]), based on query terms, indirectly constantly improve the tools that use the results themselves.

2.2. Assistive Tools for Document Selection

In this section we present the tools themselves, outlining the key findings and capabilities of each one. We also present tools which contain common attributes clustered into common themes.

2.2.1. ThemeScapes

Wise et al, implement a visualization technique by employing spatial representations of large document sets [14]. Their aim was to create a visualization that may then be visually browsed and analysed in ways that avoid language processing and that reduce

the analysts' mental load". In their research, they used Themescapes (See Figure 3) "abstract, three dimensional landscapes of information that are constructed from document corpora" and Galaxies "displaying cluster and document interrelatedness" to present the notion of document similarity. Although there was no formal user tests reported on the work provides insights of 'analysts' using the tool giving feedback of reduced time spent looking for relevant material. The users also report using the tool not just for document discovery but also for identifying document relationships (even if this is not a primary function of Themescapes).

2.2.2. VKB

Another interpretation of a collected body of materials is presented by Marshall et al [20]. In this research, a spatial hypertext tool is presented which allows information seekers to interpret results from documents and identify the structure of the document set. This is made feasible by the creation of objects, composites and collections, and allowing relationships to be defined. Building upon this early work, Shipman et al, created the Visual Knowledge Builder (VKB – See Figure 3) [15, 16, 17]. VKB supports the "incremental visual interpretation of information". This tool was thoroughly utilised for collaborative efforts on shared information space. Similarly, a prototype tool called SketchTrieve, was also created to assist information and document triage [18]. SketchTrieve, was based on a conceptual model which followed the pattern: select the services you need, connect them, press Run, and results will be displayed.

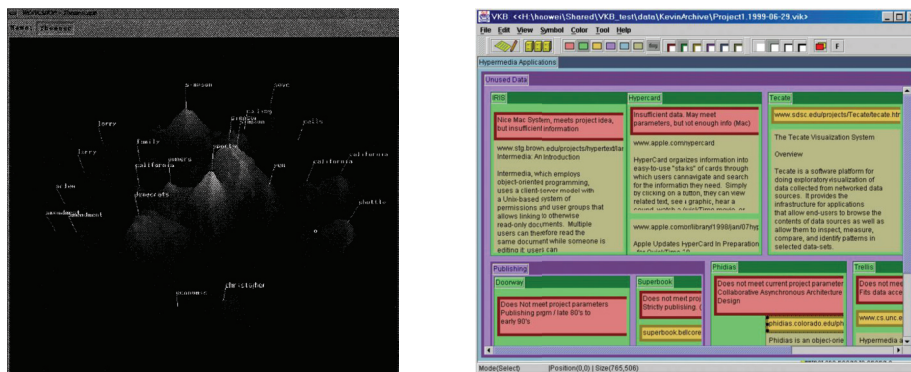


Figure 3. (Left) Themescapes and (Right) Visual Knowledge Builder Tool (VKB)

2.2.3. nSpace and TRIST

Another information visualization tool, created specifically for information triage, is TRIST (The Rapid Information Scanning Tool) [19] (See Figure 4). TRIST is built on the analytical environment nSpace [20] and allows the "rapid scanning over thousands of search results in one display, and includes multiple linked dimensions for result characterization and correlation". TRIST allows for the information seeker to compare queries and find documents that are more tailored to their need. By doing this,

document triage is informed by information that would have otherwise have taken multiple steps to achieve, all within one environment. Matching query terms to document content, like TRIST's attempt is important for information seekers. It helps them to relate their need to potentially relevant parts within a document. It is often hard however to locate the areas of the document which contain the query terms expressed by the user. A search engine will usually utilise the query terms in an information retrieval algorithm. Beyond that, there is usually no connection for the user, between the terms typed and the documents presented. Some users will use the Ctrl-F feature to find their terms within a document, but this is rarely the case [21]. It is evident that a more effective way to communicate the system's relevance decisions to the information seeker is needed. One way is to match up the query terms to areas within the documents.

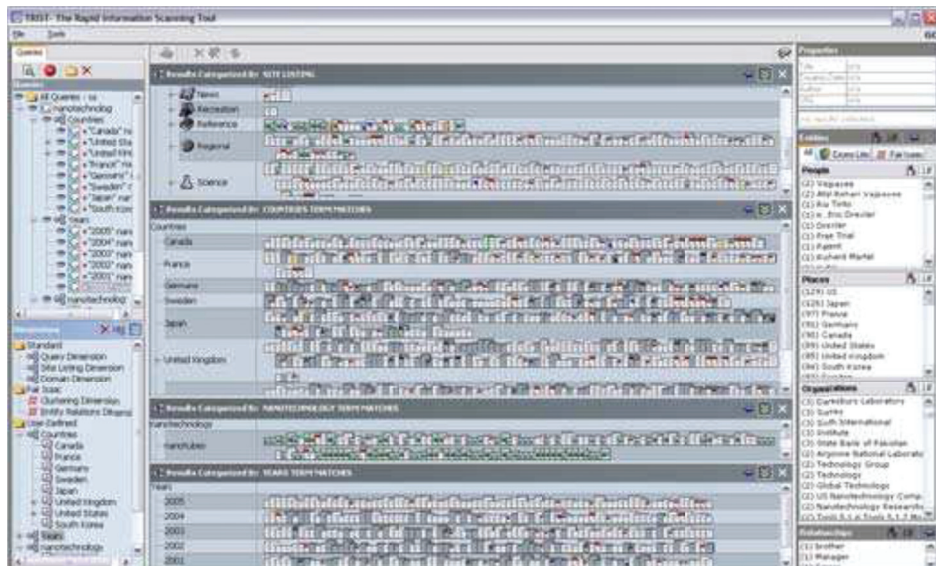


Figure 4. TRIST: The Rapid Information Scanning Tool

2.2.4. Opportunistic Search Tools

Currently, query terms are the established means by which an information seeker can make a request to a search engine. Directed browsing strategies can be assisted by several methods explained above using these terms, or variations of these, formulated by the user. Opportunistic search however, is also a big part of the information seeking process. It requires the triage of information in a less structured way. As it is becoming evident that “keyword and hypertext cannot support all these new tasks well” more opportunistic and exploratory systems are being researched [22]. One such software tool uses Semantic fisheye views (SFEV's) to browser over collections with different metrics [23, 24] (See Figure 5). A similar approach was also implemented by Cockburn et al, this time, using space filling thumbnails with a zooming action to allow better space real estate [25]. Screen real estate is one of the limitations that challenge the above prototypes. The question asked by Bae et al was whether different display types,

would have an effect on the way users perform document triage [15]. In their findings they report how there are more transitions using multiple displays rather than a single display “Additionally, users evaluated documents more by reading their contents and less often relied solely on metadata. Users spent more time reading and interacting with documents that they valued”. This corresponds with the finding that reading time correlates with assessing document value in the triage field [26].

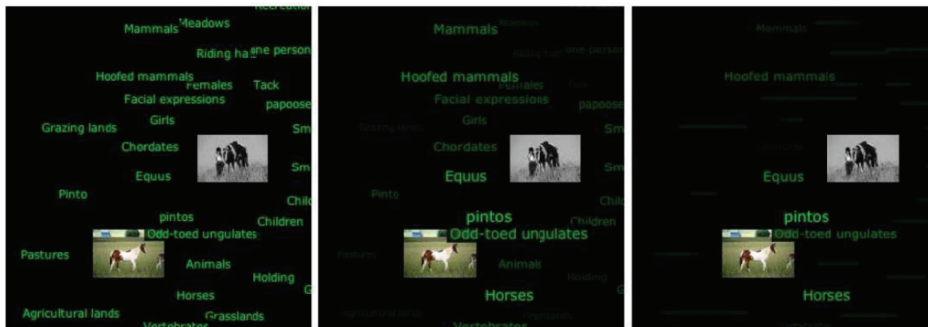


Figure 5. Semantic Fisheye Views

2.2.5. Using Task Bars

A common approach to supporting users’ triage activities is by enriched visual interfaces using scroll bars (or any bars representing the document length). Two software tools, FindSkim and ProfileSkim (See Figure 6), created visualization in the task bars, to indicate the location of query terms in a document [27, 28, 29]. ProfileSkim, also added bar charts to allow the user to find heavily populated areas, query terms wise, within the document. A similar basis was used by Donald Byrd, who used colour and term highlighting scrollbars [30] and Schwartz et al who used term distribution visualizations [31]. The argument for making use of text structure when retrieving from full text documents, has also been investigated by Marti A. Hearst, and a prototype “a visualization paradigm, called TileBars” is presented to aid the information seeker [32]. The same information and principle as Harper et al was implemented with some additions, such as snippets for reading the results before navigating towards the related area. This method was favourable with participants. Query term matching has also been used in SmartFind (See Figure 6), another hybrid Ctrl/Cmd-F tool which uses Term Frequency x Inverse Document Frequency (TFIDF) algorithms within a document to provide potentially significant document areas to the information seeker [21].

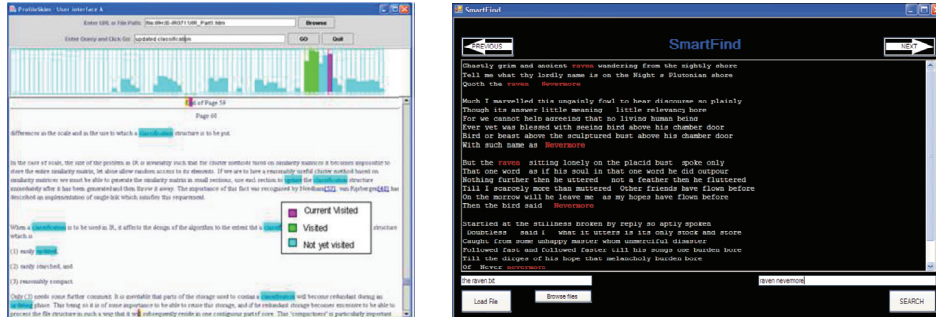


Figure 6. Profile Skim 2 Tool (left) and the SmartFind Tool (right)

2.2.6. TriDoc

TriDoc is a bespoke document triage tool which combines the high level results list view of document results with within-document scanning and information searching [33]. Currently, there are two interfaces supported by TriDoc. Both prototypes are hosted in a single-screen interface that integrates surrogate as well as within-document views; as well as snippets of individual sections of the document, combined with a full-text reading pane (See Figures 7 and 8). This approach follows the research not on visualization presentation but on the visual attention of users; a bottom approach unlike the other presented tools. The interface allows for a ‘natural’ linear type scanning of the document contents to happen in a non-linear fashion and minimizes scrolling.

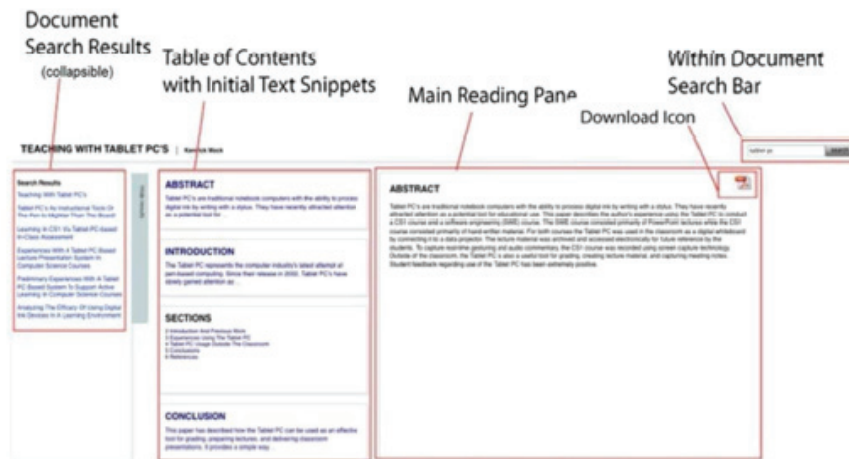


Figure 7. TriDoc Interface 1

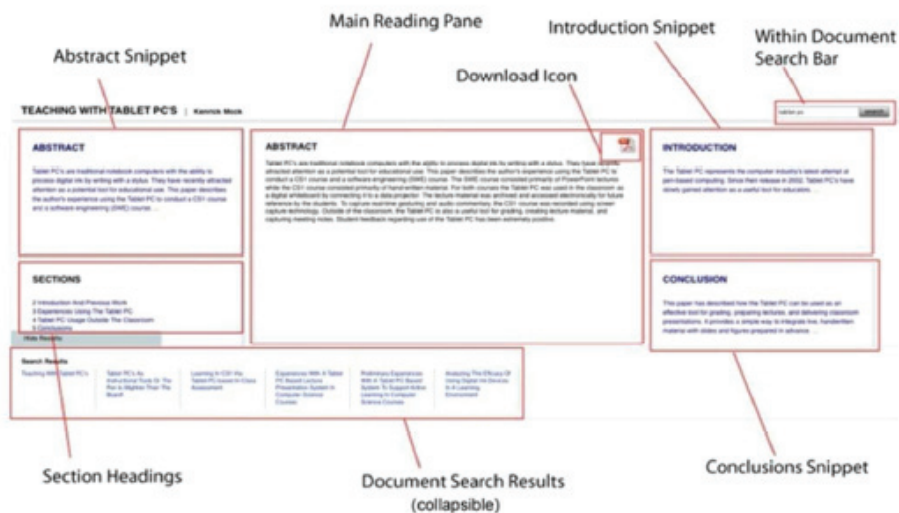


Figure 8. TriDoc Interface 2

TriDoc accommodated faster triage from the users test and received positive feedback from the information seekers (note: one interfaces received a higher rating than the other). TriDoc is an ongoing project which is currently under development (2015), unlike many of the reported tools in this paper.

3. Conclusions and Future Directions

After looking at the individual key tools which were created specifically for document selection / triage purposes we are able to make some inferences in terms of their goals, their similarities and differences. We can begin to deduce the effect on user behavior that these tools have and therefore begin to produce guidelines and understandings for designers and developers on these, similar to those by Mavri et al [34]. We were also able to comment on the implementation of these tools within commercial repositories for electronic publications.

Most tools reported on for supporting document triage use a visualization approach to present representations of the information, specifically at the highest level of triage; namely, that of the surrogate view. Interestingly, we note that very few of the tools we report on consider the individual document feature important to the users. Furthermore, each tool focuses on matching the search terms inputted by the user rather than providing representations such as document structure. This denotes a reliability on the information retrieval engine, sometimes at the cost of the manual process which occurs after by the information seeker. While, from the reported data, there is a clear improvement regarding triage performance measurements, such as time or accuracy in locating relevant information, we recognize room for further improvement at the post automatic retrieval and presentation stage. There has thus far been limited research into the actual visual attention and processes of information seekers performing within-document triage; the second stage of triage process. Furthermore, most of the findings were taken from subjective feedback rather than empirical quantitative findings. Using

the research as a theoretical foundation, we encourage tools which give emphasis on the visual attention.

Academic document searching has, until now, not been given enough scrutiny in terms of interactive interfaces for document triage in the professional field of academics. We encourage and aim to produce future work which aims to build up knowledge on this topic through the interactive behaviors through an iterative user centered design approach, rather than a waterfall model for development of the tools. A primary goal is to set a foundation for standardising the creation and evaluation of such interfaces.

References

- [1] G. Buchanan, F. Loizides, Investigating document triage on paper and electronic media, In *Procs. of the European Conf. on Research and Advanced Technology for Digital Libraries*, **35** (2007), 416–427.
- [2] F. Loizides, G. Buchanan, Towards a Framework for Human (Manual) Information Retrieval, In *Multidisciplinary Information Retrieval*, Springer, Berlin Heidelberg, 2013, 87–98.
- [3] A. Adams and A. Blandford, Digital libraries in academia: Challenges and changes, In *Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology*, 2002, 392–403.
- [4] A. Adams and A. Blandford, Digital libraries' support for the user's 'information journey', In *Proceedings of the 5th ACM IEEE-CS joint conference on Digital libraries*, 2005, 160–169.
- [5] P. R. Keller and M. M. Keller, *Visual Cues: Practical Data Visualization*, IEEE Computer Society Press, Los Alamitos, CA, USA, 1994.
- [6] G. G. Robertson, S. K. Card, and J. D. Mackinlay, Information visualization using 3d interactive animation, *Communications of the ACM* **36**(4) (1993), 57–71.
- [7] R. Spence, *Information Visualization: Design for Interaction* (2nd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.
- [8] E. Tufte, *Envisioning information*, Graphics Press, Cheshire, CT, USA, 1990.
- [9] R. R. Korfhage. To see, or not to see, is that the query? In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, ACM, New York, NY, USA, 1991, 134–141.
- [10] A. Spoerri, Infocrystal: a visual tool for information retrieval & management, In *Proceedings of the second international conference on Information and knowledge management*, CIKM '93, ACM, New York, NY, USA, 1993, 11–20.
- [11] C. Chen, Visual spatial thinking in digital libraries – top ten problems, In *Joint Conference in Digital Libraries*, 2001.
- [12] C. Buckley, G. Salton, J. Allan, and A. Singhal, Automatic Query Expansion Using SMART: TREC 3, In *Third Text REtrieval Conference (TREC-3)*, 1994, 69–80.
- [13] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, 2005
- [14] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, Visualizing the non-visual: spatial analysis and interaction with information from text documents, In *Proceedings of the 1995 IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, USA, 1995, 51–58.
- [15] F. Shipman, R. Airhart, H. Hsieh, P. Maloor, J. M. Moore, and D. Shah, Visual and spatial communication and task organization using the visual knowledge builder, In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '01, ACM, New York, NY, USA, 2001, 260–269.
- [16] F. Shipman, J. M. Moore, P. Maloor, H. Hsieh, and R. Akkapeddi, Semantics happen: knowledge building in spatial hypertext, In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*, HYPERTEXT '02, ACM, New York, NY, USA, 2002, 25-34.
- [17] F. M. Shipman, III, H. Hsieh, P. Maloor, and J. M. Moore, The visual knowledge builder: a second generation spatial hypertext, In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia*, HYPERTEXT '01, ACM, New York, NY, USA, 2001, 113-122.
- [18] D. G. Hendry and D. J. Harper, An informal information-seeking environment, *J. Am. Soc. Inf. Sci.* **48** (1997), 1036–1048.

- [19] D. Jonker, D. Schroh, B. Wright, P. Proulx, and B. Cort, Information triage with trist, In *Conference on Intelligence Analysis*, 2005.
- [20] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort, *Advances in nSpace The Sandbox for Analysis*. McLean, VA, 2005.
- [21] F. Loizides and G. Buchanan, The myth of find: user behaviour and attitudes towards the basic search feature. In *Joint Conference on Digital Libraries*, 2008, 48–51.
- [22] D. Bryan and A. Gershman, Opportunistic exploration of large consumer productspaces. In *Proceedings of the 1st ACM conference on Electronic commerce*, EC '99, 1999, 41–47, New York, NY, USA, ACM.
- [23] P. Janecek and P. Pu, Opportunistic search with semantic fisheye views. In *Web Information Systems WISE 2004*, volume 3306, Springer Berlin / Heidelberg, 2004, 668–680.
- [24] P. Janecek and P. Pu, An evaluation of semantic fisheye views for opportunistic search in an annotated image collection, *International Journal on Digital Libraries* 5 (2005), 42–56.
- [25] A. Cockburn, C. Gutwin, and Jason Alexander, Faster document navigation with spacefilling thumbnails. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, 2006, 1–10, New York, NY, USA, ACM.
- [26] P. K. Chan, *A non-invasive learning approach to building web user profiles*, 1999.
- [27] D. J. Harper, S. Coulthard, and Sun Yixing, A language modelling approach to relevance profiling for document browsing, In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, JCDL '02, New York, NY, USA, ACM, 2002, 76–83.
- [28] D. J. Harper, I. Koychev, and Y. Sun, Query-based document skimming: a user-centred evaluation of relevance profiling, In *Proceedings of the 25th European conference on IR research*, ECIR'03, 200 377-392, Berlin, Heidelberg, Springer-Verlag.
- [29] D. J. Harper, I. Koychev, Y. Sun, and I. Pirie, Within-document retrieval: A user-centred evaluation of relevance profiling. *Information Retrieval* 7 (2004), 265–290.
- [30] D. Byrd. A scrollbar-based visualization for document navigation. In *Proceedings of the fourth ACM conference on Digital libraries*, DL '99, New York, NY, USA, ACM, 1999, 122–129.
- [31] M. Schwartz, C. Hash, and L. M. Liebrock, Term distribution visualizations with focus+context, In *Proceedings of the 2009 ACM symposium on Applied Computing*, 2009, 1792–1799.
- [32] M. A. Hearst, Tilebars: visualization of term distribution information in full text information access, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1995, 59–66.
- [33] F. Loizides, Thomas Photiades, Aekaterini Mavri, and Panayiotis Zaphiris, On Interactive Interfaces for Semi-Structured Academic Document Seeking and Relevance Decision Making. *New Rev. Inf. Networking* 19(2) (2014), 67–95.
- [34] A. Mavri, F. Loizides, T. Photiadis, and P. Zaphiris, We have the content... now what?, *Information Design Journal* 20(3) (2013), 247–265.