ISSN 1744-1986

Technical Report N° 2007/ 22

Relation of Local Documents to Browsed Web Pages

Y Jacques

29 September, 2007

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

Relation of Local Documents to Browsed Web Pages

A dissertation submitted in partial fulfilment
of the requirements for the Open University's
Master of Science Degree in Software Development.

Yves Jaques
(W4913399)

March 5, 2008

14792 words

Preface

Along with all those who downloaded and used *Context Explorer*, I would like to acknowledge my wonderful test group: Claudio Baldassarre, Luca Limongelli, Gauri Salokhe, Tina Farmer, Virginie Viollier, Isabella Pieroni, Aureliano Gentile, Sylvain Caillot, Tony Jarrett, Sienna Reid, Marcel Reid-Jaques, Alan Lowther, Caterina Caracciolo, Andy Bagdanov, Elena Irde, Deb Piscitelli, Boris Lauser and Margherita Sini.

Special thanks to Alan Lowther for assistance on research methods and analysis.

Thanks to my supervisor Karl Jones who often gently steered me back on to the right track when I was wandering, helped me straighten out the mid-Atlantic mongrel, and helped me think of the project as a story.

On the software side, I am eternally grateful to Til Schneider, creator of Lucene-based desktop indexer Regain, without which my browser plug-in Context Explorer would never have been finished in time. I'm also grateful to all the other GNU products I leveraged for my application: MySql, PHP, MySqlAdmin, PHPesp, WordPress, Lucene, Tomcat, Mozilla XUL, Mozilla Firefox and MochiKit.

Lastly, my family, for gamely listening to my abstract comments about my thesis, and for bringing me plates of food, cups of tea and slices of cake while I goggled numbly at my enormous tables of logging data.

Table of Contents

List of Figures

Abstract

The relation between browsed web pages and local documents is of increasing interest as the browser takes centre stage and web indexes grow into the billions. Users are storing and accessing their data both locally and online. The division between what is remote and what is local is breaking down and yet a huge gulf remains. There is a need for research on self-adapting interfaces that contextualise the user experience to reduce system complexity. Part of this larger task lies in reducing the distance between remote and local resources.

The author developed a browser plug-in that presented users with local documents associated to browsed web pages to test the nature of the relation between remote and local resources. Users were profiled and tested in various contexts, performing work, research, search and personal tasks. They examined the relations between the pages they were browsing, their characterization as key phrases and the documents on their computers, rating the relevance and utility of contextualised local resources in a self-adapting interface.

The application also logged user actions over a period of half a year as they used or ran the application in the background while browsing, providing a vital source of data about the relation between the web and desktop computers.

Though the user group was too small to be considered statistically relevant, preliminary findings revealed that users indicated high relevance between local and remote resources when performing work, research and search tasks, and that contextualised interfaces could reduce cognitive load, minimizing the effort in finding local resources related to browsing activity.

1 Introduction

1.1 Background to the research

The computer is the centre of information distribution in modern households. It has swept away bookshelves, recipe cards and CDs. This functionality creep requires the support of a growing set of applications, files and interfaces. As complexity increases, user types multiply (Pew, 2006). The entire globe is going online, and grandparents are competing with their grandchildren for access to the instant messenger. Many of these new user types show a distinct need, not for greater complexity but for simplicity, as usability studies show the difficulties some users face with even the lean interface of the Google search page (Aula & Käki, 2005).

New research is looking at filtering the computer's endless functionality to meet user needs in the moment. *Context* has become a new computing paradigm, an umbrella concept that carries a host of buzzwords all starting with *semantic*: *semantic desktop*, *semantic web* and *semantic file system*. Semantics are seen as a way of overcoming the barriers put up by applications and file structures that compartmentalise information. Today, "users find themselves fighting with the application, struggling to simultaneously display all the information they need…[which]…is spread out over multiple applications" (Bakshi & Karger, 2005).

Numerous conferences now focus on issues of context and semantics. Ontology Web Language [1] (OWL), the bedrock of the semantic web, is now a World Wide Web Consortium recommendation. It is hoped that ontologies based on OWL will help "automated processes…to interpret their meaning flexibly yet unambiguously" (Horrocks et al., 2003).

Operating system (OS) designers have also caught the fever. Every major operating system is proposing a semantic file system that goes beyond traditional hierarchical

[1] http://www.w3.org/TR/owl-features/

representations of resources, e.g. Microsoft with *WinFS* (Code Name WinFS, 2004) and Apple with *Smart Folders*, a grafting of relational concepts over a standard storage system fronted by the *SpotLight* [2] index. Perhaps the most interesting work is being done in Linux; the still-young *GLSCube* initiative with its open-source licensing, plug-in architecture and easily exploitable application programming interface (API) (Salama et al., 2006).

At application level, the semantic desktop movement hopes to replace users' current working practices by giving them new, integrated personal information management systems. Massachusetts Institute of Technology's *Haystack* client is one of the more mature, with a full-featured personal information manager (PIM) that promises to "let users define their most effective arrangements and connections between views of information" (Haystack Project, n.d.). *Chandler* is another example, "organized around your data and the semantics you attach to it...not around technologies and features" (Chandler, n.d.).

In the same period that the semantic desktop movement has placed its emphasis on developing new applications the browser has gone on a meteoric rise, becoming the centrepiece of virtual social collaboration. Activities span research offerings such as *OntoWiki* [3] to significant industry investments such as *GoogleDocs*[4] and Microsoft's collaboration server *SharePoint*.[5] By now it is clear that any contextual solution that does not place the browser centre stage is bound for the trash heap. The browser is no longer just for viewing web pages. It has gained (or suffered, depending on your viewpoint) from requirements creep: becoming an email client, a news reader, a document and spreadsheet editor, a games console and an image gallery. Every month sees a new application ported to the browser. These new applications are also increasingly sophisticated as bandwidths grow and techniques such as Asynchronous Java and XML (AJAX) extend the limits of browser functionality (L. Paulson, 2004). Several interesting browser plug-ins are experimenting with contextual concepts, such as ClearForest's

---

[2] http://www.apple.com/macosx/features/spotlight

[3] http://ontowiki.net/Projects/OntoWiki

[4] http://docs.google.com/support/

[5] http://office.microsoft.com/sharepointserver/

Gnosis, which offers "real time semantic processing for the web…evaluating the pages you read – as you read them" (ClearForest, n.d.).

At this point the name *browser* has a decidedly antique ring to it. *Querier* would be more appropriate given that query-based searching has trumped hierarchical browsing as the preferred way for users to find relevant resources. Anyone who doubts this needs only look at the world's most visited home page, Yahoo!, where the hierarchy of subject areas that long dominated this founding search engine's page is now gone.

If the browser has become the *desktop* environment for most users, in order for it to take its true role at centre stage it must preside over a marriage of applications and resources. As applications are increasingly ported to the browser, the need for it to become a contextual virtual desktop is apparent. We need a browser that bridges the gap between the network and the local system; one that uses contextual techniques to filter and organise the user's vision of resources, applications and tasks. Today, this vision is fragmented across systems and networks and is far too static to be considered user-friendly. Users need systems that are adaptable, reconfiguring themselves based on profile, behaviour and active resources.

1.2 Aims and objectives of the research project

The aim of the project was to examine and compare computer and human-centred views of the gap between external web pages and local documents. Can contextual self-adapting interfaces act as a bridge by dynamically filtering local resources using information retrieval techniques such as key phrase extraction and document indexing? As applications are ported to the browser and the storage of emails and documents is increasingly moved to external servers, local system boundaries are breaking down, yet integration between local and external resources remains elusive.

The project objectives were to:

- produce a literature review that discusses the state of the art in human and computer-centred approaches to resource discovery, including interface design and existing application solutions;

- pose testable hypotheses about the utility of contextualising remote and local resources;
- create a browser plug-in that could be used to test the project's hypotheses;
- create a centralised logging system to store user data for the project's hypotheses;
- create a user survey to characterise the user group;
- create task scripts and questionnaires for user observation;
- test a group of users using the application, user survey and task scripts and
- support or dismiss the hypotheses using the logging and user observation data.

1.3 Overview of the dissertation

This dissertation consists of a literature review, research question and hypotheses, research methods and techniques, data collection, results analysis, validation and conclusion. Appendices contain user testing materials, application screenshots and the logging data model. Software code and analysis data are included on an attached CD.

The literature review discusses the current body of knowledge on information context and relevance. Computer and human-centred approaches are analysed, as are user interface design, personal information managers and desktop indexers. The review also contains the research question and a set of hypotheses.

This is followed by a section on research methods covering the setting, data types, measures attempted and why those choices were made based on the research aims and study design.

The dissertation then covers data collection: what data was collected and how and the limitations and problems encountered during collection. Included is a discussion of the steps taken to pre-process the data and prepare it for analysis.

Data collection then leads to the results. The hypotheses are answered one-by-one using the collected data and a look is given at how well the results generally answer the research question. Finally, there is the consideration of whether the analysis methods and design were appropriate and how they could be improved.

The dissertation concludes with an overview of the project: the objectives completed, the broad outcomes and how future work could continue to expand upon the research area.

2 Literature review

2.1 Current body of knowledge

While computer-centred approaches have long dominated the study of *information retrieval* (IR), human-centred approaches based on the study of *information behaviour* have become the focus of new studies. This shift is a development long predicted by leading IR researchers such as R. Baeza-Yates and Ribeiro-Neto (1999), who note in the landmark *Modern Information Retrieval* that computer-centred approaches to IR, although "dominant in the market place…[are]…unlikely to yield a good solution to the information retrieval problem in the long run". A fundamental oddity exists in that indexes provide a view of retrieval that is system-based and often not correlated to actual user needs. Information behaviour attempts to fill this gap between users and systems. Whereas computer-centred approaches highlight information and how to "structure, store and retrieve it", human-centred approaches focus on "trying to understand how people interpret and use information" (ibid.).

The marriage of these approaches to enhancing user experience is an area of research that is gaining popularity, particularly in the realm of web and social/search collaboration. Baeza-Yates and Ribeiro-Neto (ibid.) mention the possibility of combining IR and browsing, noting that it is "not yet a well-established approach and is not the dominant paradigm…[but]…while combining information and data retrieval with browsing is not yet a common practice, it might become so in the future".

2.1.1 Computer-centred approaches

Computer-centred approaches to IR deal primarily with statistical and semantic entity correlations between documents. These approaches take a narrow view of information behaviour, focusing on cognitive frameworks such as Belew's *Finding out about*, which comprises "research activities that allow a decision-maker to draw on others' knowledge" (Belew, 2000). He considers the process to be based on question–answer cycles: 1) asking a question; 2) constructing an answer; 3) assessing the answer. Belew does, however, consider it to be a dialog – an iterative process that works towards a successful result however constrained it may be to the result set retrieved by the queried information system.

The same approach is discussed by Baeza-Yates and Ribeiro-Neto (1999). The process is seen to be system- not user-based. The first step is that a "logical view of the documents is defined" and an index created. It is only at this point that a user enters with a *user need*, which is "parsed and transformed". *Query operations* are applied to produce a *query* that is a "system representation for the user need". This query is processed and obtains *retrieved documents*, which are first "ranked according to *likelihood* of relevance" before being returned to the user. The user might then begin a *user feedback* cycle.

Relevance in the computer-centred approach is algorithmically based, using mathematical rules to determine the relatedness of a user query with a resource. The IR model used determines the prediction of relevancy. A variety of models based on set, vector and probabilistic methods exist, of which only a few have survived the gruelling test of marketplace survival. Classic Boolean models with rigid set theory have all but vanished, with variations of set, vector and probabilistic models still left in the ring.

By far the most common model today both for its elegant term-weighting and efficient performance is the *vector-space model*, which considers every key phrase in a document to be a vector in a multi-dimensional space. The coordinate of a document in the direction corresponding to a term is a function of *term frequency* (how often a term appears) and *inverse document frequency* (scaling down the weight of terms appearing in a number of documents). The final proximity of a document to a query is also adjusted for document length, as long documents are naturally further from queries than short ones are (Chakrabarti, 2003).

Moving from models to the representation of documents, what Baeza-Yates and Ribeiro-Neto term the *logical view* of a document is a process of extracting key phrases. The steps employed in arriving at this logical view are 1) lexical analysis to treat digits, hyphens, punctuation and case, 2) the elimination of stop-words such as *the*, *a* and *an*, 3) the stemming of remaining words to remove affixes and syntactic variation, 4) the selection of index terms or term groups (nouns typically carry more weight) and 5) the construction of term categorisation structures, usually using thesauri (Baeza-Yates & Ribeiro-Neto, 1999). These steps produce the logical view of the document that is actually indexed and used for querying.

In an attempt to overcome the limitations in characterising documents by their indexed terms, several interesting lines of research exist that begin to bring computer-centred approaches closer to their human-centred cousins. Probabilistic methods rely on a seed set of results marked by users as relevant, in order to begin returning results (ibid.). *Latent semantic indexing* (LSI) characterises both the documents and the queries by moving them into their concept space (Dumais, 1988) to take advantage of known relations among concepts.

The problem with semantic approaches in general is not only that they are as yet "unscalable in terms of both memory consumption and computation time" (Tang et al., 2004), but that there are additional troubles with the semantics themselves. It is more difficult than initially thought to move objects into a shared conceptual space. "When the corpus is large and heterogeneous, LSI's retrieval quality is inferior" (ibid.). Ontology engineers have developed numerous ontologies in an attempt to characterise conceptual relations. These approaches are successful usually within a well-defined domain that has tightly agreed-upon definitions; e.g. scientific taxonomies. Concept spaces appear tightly linked to user groups with narrow interests. Nevertheless, ontological approaches still put the focus on the system and its contents, not on the analysis of user need.

2.1.2 Human-centred approaches

Information behaviour is a broad branch of study that incorporates aspects of library science, psychology, sociology, marketing and even health research to try and understand how humans interact with the information around them. Of greater interest to IR is information behaviour's sub-branch *information-seeking behaviour*, defined by pioneer researcher T.D. Wilson as the "purposive seeking for information as a consequence of a need to satisfy some goal" (Wilson, 2000).

The shift from computer-centred to human-centred approaches has meant a social sciences research shift from quantitative to qualitative measures. Opposed to the classic Text Retrieval Conference goals of quantifiable measures for speed, recall and precision, primary methods are often observation and semi-structured questionnaires. To accompany this shift in method, several models of information-seeking behaviour have been proposed. Wilson's model looks at information-seeking behaviour as the attempted resolution of physiological, cognitive or affective needs that lead to demands (often iterative) on information systems and sources, or set up an information exchange with other people (Wilson, 1999). The model also considers the user's role in the situation, such as *work* or *personal*, which he terms "context". In Wilson's view, the user encounters information-access barriers arising from these contexts. Wilson's work was later expanded by combination with a category of user action stages first proposed by Ellis: starting, chaining, browsing, differentiating, monitoring, extracting, verifying and ending (Ellis, 1987). Models such as these provide a framework from which hypotheses about human-centred approaches to IR can begin to be tested.

A model that goes further than Wilson's is Dervin's *sense-making* theory. On one side of the act of information seeking, it focuses on the user's current *situation*, and on the other it looks at the desired *outcome*. In between these two states, the model contains the *gap*, posited as the difference between the current situation and the desired situation, and the *bridge*, which is the means by which the user crosses that gap (Dervin, 2003).

Sense-making is employed as a cornerstone of user testing in the field of HCI. Its focus on *situation–gap–outcome* is useful in analysing user actions in scripted task observation sessions. It lends itself well to what Wilson terms *information-searching behaviour*, the "'micro-level' of behaviour employed by the searcher in interacting with information systems of all kinds" (Wilson, 1999).

The rise of the web and its hyperlinked structure has changed how information-seeking behaviour is viewed by researchers. Once largely ignored by IR studies, activities such as *navigation* and *browsing* have come to play an important role in the search process. "The web has legitimized browsing strategies that depend on selection, navigation and trial-and-error tactics" (Marchionini, 2006).

As searching moves into an ever wider sphere of activities, human-centred approaches begin to shed light on the differing needs and behaviours of user groups. Instead of focusing on the system, they focus on users in an attempt to typify groups, group behaviours and activity types. It is this research that has shed light on the changing world of information seeking and demonstrated that users often move fluidly and unconsciously between querying, navigating and browsing, a realisation that has led to new models of user behaviour.

One of the first to challenge the early human-centred models focusing on information-searching behaviour (Wilson's micro-level) was the interface-oriented Bates and her *berry-picking* model. This model looks at the shifting behaviour of users within the search process (Bates, 1989). Taking a phenomenological approach, Bates insists that the process of seeking modifies the seeking – a view that, while once revolutionary, now seems obvious given the haphazard way many users satisfy their information needs by wandering across queries, pages and sites while popup windows explode across their screens. Indeed, Bates' model makes the point that users are not trying to produce the perfect result set but instead are gathering bits (berries) along the way.

The berry-picking model takes it as given that relevance feedback is an essential part of the process, and it is here that we find the sacred point of communion between human-centred and computer-centred approaches. Relevance feedback remains one of the final frontiers for computer-centred approaches. Having more-or-less maximised the possibilities in providing algorithmic relevance, IR has at last turned its gaze to the user.

Explicit relevance feedback attempts to learn from user indications of which results are actually relevant in order to either expand or refine underlying queries with query re-weighting (Baeza-Yates & Ribeiro-Neto, 1999). Early attempts were largely unsuccessful owing to user dislike of the effort involved in affirming or denying results. Newer approaches are proving more successful, such as Shneiderman's dynamic query interfaces that use more advanced HCI techniques allowing for the zooming and brushing of various parts of the result set to permit quick previews (Voorhees, 2005). A closely related strategy is *boosting*, which takes user-indicated relevance to automatically re-sort, not re-query, the current result set. *Blind* or *pseudo-relevance* feedback is a computer-centred strategy that boosts items in the result set that share statistical or semantic similarity to highly relevant machine-ranked results.

Implicit feedback is yet another strand that interprets user actions to assign relevance. A promising area of research within implicit feedback is *social search*, which weights relevant items across user communities or other groups of users based on their aggregate choices. A notable search engine using such technology is *Eurekster*.[6] Another example of a social approach to relevance is *Del.icio.us*,[7] with its social bookmarks approach.

The area of implicit feedback has seen frenetic activity in the last five years. As internet advertising revenues increase, the importance of monitoring and learning from group and single-user behaviour has grown tremendously. Systems that employ these techniques are known as *recommenders* or *intelligent personal assistants* (Chaudhri, 2006). The various methods in use include session profiling, in which a user's session with a particular web site is logged. These logs are leveraged to anticipate future user needs. Large internet-based retail sites such as Amazon make heavy use of this technique to suggest related items. Some groups are experimenting with so-called *semantic periods*, in which user internet activity is logged in periods such as 'key phrase search', 'viewing' or 'browsing' in an attempt to more usefully profile behaviour (Morita et al., 2006). The profiling of actions from a user's local machine is another example of behavioural monitoring. This kind of profiling can monitor user behaviour across sites and applications, attempting to anticipate future needs from past actions.

Within the more restricted domain of querying, these techniques are used to boost the relevance of selected resources or of resources related to previously selected resources, theoretically improving the relevance of search results over time. Google offers this optional service to users, but privacy concerns have made many shy away (Johnson, 2007). A similar approach based on a social theme is the use of group behaviour in tagging photos to construct folksonomies that can be applied to individual users when searching or tagging their own photos – a technique used by large image-gallery sites such as *Flickr*.[8]

2.1.3 User interface design

While both human- and computer-centred approaches have developed models and techniques to meet the goals and tasks of users, interface designers have kept apace, experimenting with a dizzying variety of visual possibilities. The complexity of designing

---

[6] http://www.eurekster.com/about

[7] http://del.icio.us/about/

[8] http://www.flickr.com/photos/tags/

clean search interfaces that present data clearly is a problem that will never be solved completely. Nevertheless, a number of principles have been developed.

First among the principles of search scenarios is feedback: "Information that is sent back to the user about what action has been accomplished upon the use of a control" (Stone et al., 2005). Users need to understand the relation between the query or, more broadly, the task they are performing and the results with which they are presented. IR uses a variety of techniques, the most common being the document surrogate, "consisting of the document's title and a subset of important metadata" (Baeza-Yates & Ribeiro-Neto, 1999). Increasingly common as bandwidths and processing power climb are *key phrase hits within content* (KWIC), in which phrases or sentences surrounding key phrases located in a document are displayed. Clustering visualisations instead show the relation of results to other results using categorisation and sometimes screen distance from other clustered results.

Reducing cognitive load is another important search-interface principle. One of the four psychological principles of HCI is that "users have difficulty focusing on more than one activity at a time" (Stone et al., 2005). It is easy to become lost in various search strategies while struggling towards a goal. Given Bates' berry-picking model, it is clear that many users benefit from having a basket in which to place their berries. It is now common to provide both implicit and explicit ways in which to save choices. Another technique is multiple panes, in which users receive additional details about a result without losing their overall place.

The last design principle tightly bound to search activities is that of customisable interfaces supporting various interaction styles and beginner/expert options. "Blending several interaction styles is particularly appropriate when…the experience of the users is varied (ibid.)." User abilities and needs vary, and accordingly, much work has gone into designing interfaces that can be used with little prior knowledge but are capable of being scaled to greater complexity as users advance in ability or require a larger set of search options.

A variety of visualisation techniques have been developed for browsing and query construction: icons, the colour highlighting of text, brushing and linking (the connecting of two views so that changes in one affect the other), panning and zooming (showing clusters that can be zoomed in on), focus-plus-context (fisheye), magic lenses (modifying data under the lens) and animation (to show nodes that would otherwise be hidden in trees).

For browsing, a common technique is to use categories or directories that contain wider results, or key phrases that are used to generate queries. Also common is faceted browsing, in which multiple dimensions are browsed and selected, acting to filter the result set. The remaining techniques are clusters and maps, both of which group related key phrases or results into common areas. Query construction on the other hand generally uses text input, often with Boolean operators such as 'and' to restrict result sets, while other techniques may overlap with those of browsing, e.g. facets. A remaining area is query expansion and natural language querying. Both show mixed results.

2.1.4 Existing application solutions

As interface design techniques are exhausted, as classic IR maximises its algorithms and as information behaviour models mature, the field has been searching for new directions. Into this gap the semantic desktop movement has boldly stepped. It is a mixture of IR techniques with a strong admixture of semantic indexing and an armful of user-behaviour studies and HCI principles. The movement hopes to reduce repetitive tasks and cognitive load by taking a (meta) data-centric view of users' systems to seamlessly offer data and applications that meet their goals and tasks. A strong orientation is an emphasis on personal and small-group data, an idea covered as early as 1999 by Marti A. Hearst writing that "another trend that can be anticipated is an amplified interest in organization and search over personal information collections" (Baeza-Yates & Ribeiro-Neto, 1999).

Xiao and Cruz (2005) note that the main features of such systems are "semantic data organization", a method that goes beyond hierarchical systems and allows for "finer data manipulation". Secondly, through "flexible data manipulation" these systems can perform activities of "integration, exchange, navigation and query processing" on the data they handle. Thirdly, "rich visualisation" combines what are normally separate views of data into a unified whole.

This is an ambitious goal, but the possibility is too enticing to resist. A number of research applications have been developed, but the results are not encouraging. Instead of working under the hood to make OSs and existing applications behave in a more organic fashion, many projects, e.g. those of Iris[9] and Chandler,[10] have focused on developing new applications that integrate common tasks such as emailing, making appointments and searching for resources. This requires a total re-engineering of the business process for users, most of whom are understandably unwilling to give up the applications they already use.

Recent developments by the Gnowsis team, concentrating instead on a powerful semantic-services layer, seem a step in the right direction. "Data structures are not changed and existing applications are extended and not replaced" notes the Gnowsis documentation (n.d.). Their approach offers an agnostic ontology that is personalised for every user (Fernández et al., 2006) and an API that existing applications can use to add integration, offering a step-wise improvement to end users without requiring them to abandon their cherished applications.

A laudable objective of these initiatives is improving collective workflows for group data – pipelines that have long been underserved by traditional applications. From the standpoint of information behaviour, the interesting aspect of social collaboration is that, when classic IR fails users, they build their own mechanisms using such means as social book marking, forums, blogs and elaborate folksonomies. This lesson must not be lost by the IR community if it is truly to gain from the prospect advanced by new browser-based applications that use AJAX to deliver rich functionality from a global resource base.

A simpler set of application solutions that can be considered to be a partial response to problems of local/remote resource retrieval are the browser toolbar and plug-in solutions offered by a number of major and minor players in the indexing and web-search field such as Google desktop[11] and Yahoo! desktop.[12] A common feature of such applications

---

[9] http://www.openiris.org/about-iris

[10] http://chandler.osafoundation.org/philosophy.php

[11] http://desktop.google.com/features.html

[12] http://beta.in.desktop.Yahoo!.com/features/

is a local index that allows users to concurrently search the web and their local systems. However, such approaches are no more than a marriage of two indexes, one local and one web-based. They do nothing to answer the behavioural aspects of the information problem, doing little to contextualise the workspace and relying on specific user-formulated queries to provide results; they do not anticipate user behaviour or needs.

Human information behaviour is a complex, multi-faceted problem that provides many opportunities for combined approaches. IR and artificial intelligence experts with statistical, probabilistic and semantic techniques, information behaviour researchers with models and studies of information-seeking behaviour, interface designers with minimalist approaches to user interfaces – each holds a piece of the puzzle. What is needed is an emphasis on shared research that brings these disciplines into a common field to build the next generation of self-adapting interfaces that from the ground up hide complexity and anticipate user needs. Indeed, the spontaneous growth of Web 2.0 with its blogs, wikis and folksonomies represents an authentic organic expression of common user need for collaboration and context in information resources.

2.2 Research question

This project answers the question: How practical and useful is the automatic association of indexed local documents to browsed web pages using the key phrase extraction of web pages to query local resources via a browser plug-in and dynamically display the results?

To provide a means of analysing *practical* and *useful*, this study gathers and analyses data to support a set of hypotheses about user behaviour during various tasks: work, research, search and personal. Using a mixture of auto-generated logging data, a survey and user observation and questionnaires, the relevance between web pages, extracted key phrases and related local documents is measured. The hypotheses are grouped into four areas: task context, user type, relevance factors and manual intervention.

2.2.1 Task context

The application is more useful for activities that are performed both locally and remotely

It is thought that the application is useful only for users performing a browsing activity linked to work they perform on their local system with local data.

Fewer results equates with lower use

Users who obtain few results are less likely to use the application.

There is a correlation between often-visited pages and a greater relevance or quantity of local file results

Pages that users often access may be more tightly related to common activities such as work tasks. It follows that these pages are more tightly related to the resources on a user's local system.

2.2.2 User type

Users who spend more time organising their data in many folders have more relevant results

Users who spend time organising their local documents in many hierarchical folder structures and/or use naming conventions may be more actively involved with their data, and this may raise the likelihood of relevant results.

For some users, folders play an important role in interpreting results

Users who spend significant effort organising documents in hierarchical folder structures will show an increased interest in viewing results as folders rather than files.

Users recognise files by their filenames

As most indexed filenames on a local system are user-generated, it follows that users may need no extra metadata to recognise result sets.

2.2.3 Relevance factors

Only the most relevant results are at all relevant

Given the fuzzy match between web page meaning and extracted key phrases, it follows that only the most relevant resources (as determined by the vector-space model) are actually relevant.

The more key phrases extracted, the less relevant the results

As pages with more content are probably broader in scope, it follows that these pages should be less relevant to the results.

Frequently occurring key phrases are associated with more relevant results
Often-occurring key phrases may indicate a user's area of interest or common work tasks and so be more closely associated with their local documents.

2.2.4 Manual intervention

Filtering is a useful selection mechanism to increase relevance
Filtering results by document type or key phrase is helpful in increasing result set relevance, particularly as result sets grow larger.

Key phrases extracted from title and metadata create more relevant results
Title and web page metadata contain a concise representation of the document meaning and should create more relevant key phrases and more relevant local results.

Query expansion using automatic relevance feedback is not very useful
Given the fuzzy match between web page meaning and extracted key phrases, it follows that query expansion using the highest-rated resources will tend to further muddy the relevance between the web page and the local file results.

2.3 Summary

Classic IR researchers admit to having largely exhausted the potential of computer-centred approaches. Human-centred researchers have focused much of their effort on understanding how users search for information and have developed a number of interesting models. HCI research has taken human-centred research and applied it to interface design to streamline and simplify the user experience.

Research that combines these techniques is the most promising road to meet user information needs. It is a strategy that mirrors the current trend of the merger of Web 2.0 and its spirit of social collaboration with the structured semantics and machine-readable content of the semantic web, a movement sometimes now hailed as Web 3.0.

In all this we must not forget that users have invested significant amounts of time in understanding how best to use their current systems and have developed deeply ingrained work practices. If we are to move them towards contextual, intuitive interfaces and shared semantic data spaces, we must implement step-wise, scaffolding methods that slowly leverage existing applications and work practices to move people gradually and instinctively towards a shared semantic future. The case study application built for this thesis attempts in its own small way to do just that, by unobtrusively closing the gap between the user's web browsing activities and the related resources on their local system.

3 Research methods

3.1 Research setting

Context Explorer,[13] a Firefox[14] browser extension application (see *Appendix G* for screenshot and *Appendix I* for source code and installer) was created that leveraged open-source desktop indexer Regain[15] and the Yahoo! Term Extractor.[16] It interacted with the browser and the local system to relate web pages with local documents.

Regain scans and indexes local-system text resources such as Microsoft Office and Adobe Acrobat documents. It stores these in a regularly updated index. Context Explorer extracts the key phrases from browsed web pages using the Yahoo! Term Extractor and builds with the results a query that uses Regain's index to find relevant local documents and display them in the browser sidebar.

Because the aim was to compare web pages and local documents, a browser was the natural testing platform. The Mozilla foundation's Firefox browser was chosen owing to its elegant *XML User Interface Language*[17] (XUL) plug-in development framework. For local document indexing, the Regain desktop indexer was chosen for the ease with which its output could be modified to meet plug-in needs and for the fact that it used *Lucene*,[18] a popular battle-tested open-source indexer that uses the classic and effective IR technique of vector-space modelling that lies at the heart of most indexing engines.

The choice of the key-phrase extraction component was tightly linked to study needs and study design. Several other key-phrase extractors were examined, but they suffered

---

[13] http://www.organicdesktop.com

[14] http://www.mozilla.com/firefox/

[15] http://regain.sourceforge.net/

[16] http://developer.yahoo.com/search/content/V1/termExtraction.html

[17] http://developer.mozilla.org/en/docs/XUL

[18] http://lucene.apache.org/java/docs/

from important defects. Most important was whether or not they were accessible as a web service. Any method of data exchange that was not URL-based implicated serious additional development work not in line with the one-year time frame of this dissertation. Because the only component that was easily exploitable as a URL-based web service was the Yahoo! Term Extractor, the choice was clear from the standpoint of utility. Secondly, most other extractors require corpus training before being able to return any key phrases, which was too great a burden for the user observation sessions. A review of the literature also showed that it was being used by other researchers in similar situations. It admittedly suffers from a flaw in that its exact algorithm is a corporate secret and is unpublished. However, to mitigate this problem, user-testing questioned not only the fit of local documents to web pages but also the fit of extracted key phrases to web pages.

3.2 Research design

3.2.1 Logging

Integrated into Context Explorer was a central server database-logging mechanism (see *Appendix H* for the data model) capable of recording all interactions with application instances.

3.2.2 Survey

To assist in the user profiling, a survey (see *Appendix B*) focusing on computer literacy was employed. A transcription of Ulla Bunz's (2004) *Computer-Email-Web (CEW) Fluency Scale*, it allowed for comparisons between the Context Explorer user group and the user group profiled by Ulla Bunz.

3.2.3 User observation

Other than quantitative user data and user-profiling via survey, the other main area of research was qualitative action research. An evaluation strategy was defined using semi-structured field observation based on task scripts (Stone et al., 2005). It consisted of four scripted browsing tasks (see appendices *C* through *F* for observation, task and training scripts and the logbook template) using think-out-loud protocols (Open University, 2001). Each task contained a set of subtasks with questions about the process. The testing materials were piloted through three iterations until the data collection satisfied study needs without requiring too much effort from the volunteer group.

3.3 Research Techniques

3.3.1 Primary and secondary data

The research based on the case study application involved the collection of primary data gathered through observation (Sharp et al., 2002): user logging, survey, user observation and questionnaires. As the project was a case study based on a new application of which other examples were not found, it was impossible to compare this primary data with secondary data from similar applications.

One area nevertheless allowed for a comparison of the primary data with existing secondary data; the user group involved in the study was profiled using an existing survey on web fluency and compared to other groups.

3.3.2 Quantitative and qualitative measures

The study had an overarching goal of measuring the task-dependent relevance of the result sets of local documents to browsed web pages and the overall utility of the application as perceived by users. Relevance and utility were measured both quantitatively and qualitatively. There was a measurable quantitative aspect using numerical measures of the precision of query result sets, i.e. the ratio of relevant results to returned results (Clarke & Willet, 1997) for results above a one-percent threshold of Lucene-reported relevance (this cut-off was necessary, as otherwise, users would have had to analyse every result, and sets sometimes ran into thousands). There was also a qualitative observation-based aspect of how relevant the contextualisation was to the tasks that users were given to perform with the application.

Though it would be interesting to also measure recall, i.e. the ratio of relevant returned resources to overall resources (ibid.), the task was too burdensome for a volunteer user group and too time-intensive for the project. Though Clarke and Willett promote the use of a technique they term *relative recall*, it was not usable, as it relies on a comparative return from multiple search engines. This would have required the installation of multiple local search engines on user computers, which, again, was too burdensome for the volunteer user group and project timeline.

3.3.3 Data types

The logging data were at a minimum categorical and allowed the classification (Sharp et al., 2002) of browsing activity. In addition, much of it was also used to create ordinal and ranked data to identify the frequency of certain actions. Some of this data were also meaningful as interval and even ratio data, which was used to identify user behavioural patterns and draw inferences.

The user observation data were a mixture of interval (e.g. five-point scale), ratio (e.g. percentage of relevant documents) and categorical (e.g. which task is the most relevant) data.

3.3.4 Analysis techniques

Before analysing the data to understand whether an effect or correlation has been found, there are at least three aspects that must be considered: sample size (statistical power), correlation strength and correlation significance.

Correlations were calculated wherever the study looked for a trend between two variables. "Correlations are perhaps the most basic and most useful measure of association between two or more variables…[and]...provide information about the direction…and the intensity of the relationship" (Marcsyk, 2005). Of the various methods for calculating correlation, the Pearson *R* was used, as it is designed for ratio and interval scales.

The study data had two very different sample sizes. The logging data contained thousands of records, while the user observation data consisted of tens of records. Thus, while testing for significance was not as important for the logging data owing to its large sample size, it was quite important for the user observation data, to determine not only whether there was a correlation but also the significance of that correlation. "Tests of significance allow us to estimate the likelihood that a relationship between variables in a sample actually exists in the population and is not simply the result of chance" (ibid.).

Owing to the relatively small user group (19 users), the *t-test* method designed for a group of below 30 was used.

Although for the inferential statistics related to interface usability a random user group would be preferable, it was not possible within the constraints of this project. While a t-test was used for the five-point scale of "not at all" to "excellent" related to relevance scoring, for categorical measures such as whether users found one task more related than another, a chi-square was used. "T-tests…are appropriate only when the dependent variables being measured are continuous (interval or ratio). In contrast, the chi-square statistic allows us to test hypotheses using nominal or ordinal data…chi-square analysis is often used to examine between-group differences on categorical variables, such as gender, marital status, or grade level" (Marcsyk, 2005).

4 Data Collection

Data collection was a multi-pronged effort of online surveys, task scripts and questionnaires. Automatic logging data were continuously collected from July 2007 to January 2008, eventually generating data for tens of thousands of page loads by over 30 users. User observation sessions conducted in autumn 2007 were held at users' computer stations on their regular work, study or home computers.

4.1 Process and data sources

4.1.1 Context Explorer application

Context Explorer contained a variety of data-gathering features. Users could select the content extraction type: all, content, title or metadata. Each extracted key phrase showed the number of local documents found and could be used to filter results. The found document types were also displayed and could be used in filtering. A window displayed the available local files together with mouse-over pop-ups showing document metadata. Clicking allowed users to open the file or the folder in which the file was found. There was an optional folder view that showed relevant folders instead of files. A final feature was a query expansion button, which allowed users to re-query using summary text from the most relevant current documents.

4.1.2 Logging system

Quantitative user data were logged (see *Appendix I*) from each instance of Context Explorer. The system logged all visited web page URLs together with their content length to detect changed pages. All extracted key phrases were logged. The number of local resource hits returned per key phrase and the index size were also stored to allow an analysis of the relative relevance of pages and key phrases to local systems. Document types and quantities of folders were monitored. Lastly, all user actions were logged, such as browsing file metadata, refining key phrases and filtering document types. These data were stored against the user's randomly generated Firefox profile folder name, allowing the strict separation of users while also ensuring anonymity.

4.1.3 Surveys

Fourteen out of the nineteen observation session participants completed the survey online before participating in the user observation session (see *Appendix B*). Results were stored in a central database (see *Appendix I*).

4.1.4 User observation

User testing lasted for about three months and included 19 volunteers. Users were recruited from friends and co-workers, and observation sessions were held using users' computers, either at home or at their workplace.

Results of the user observation tasks were recorded on paper and transcribed to a database to facilitate analysis (see *Appendix I*).

4.2 Preliminary analysis

For the logging data, it was striking to see how many key phrases web pages have in common. The key-phrase database contained approximately 60,000 unique key phrases, while the join table between key phrases and the pages upon which they were found had about 250,000 entries, meaning that on average any one key phrase was found across about four pages.

An observation following upon this was that the database join table correlating key phrases to local resources was only about half the size – approximately 120,000 entries – meaning that about half the key phrases extracted from web pages did not produce any results on any user's computer, showing that users are generally browsing pages that are thematically wider than the local documents on their computers.

A third interesting observation was that the pages browsed by users were quite dynamic. For web pages, an entry was created every time a page was accessed and its string length had changed. Analysis showed that, for any URL, a user visited it an average of only 1.4 times before it had changed, though much of this may be owing to advertising content.

Moving from the logging database to user observation, it was clear even before analysing the data that there is much greater correlation between local and remote resources when users are performing work/study activities. The local hit counts were always higher for these tasks, as were user marks for relevance between key phrases and web pages.

5 Results

5.1 Implementation based results

The Context Explorer application was the central mechanism used to answer the research question. It dynamically displayed both the key phrases extracted from the current web page and the documents on the user's computer that were related to those key phrases. It allowed for automatic logging and user observation sessions in which the relationship between web pages, key phrases and local documents could be explored.

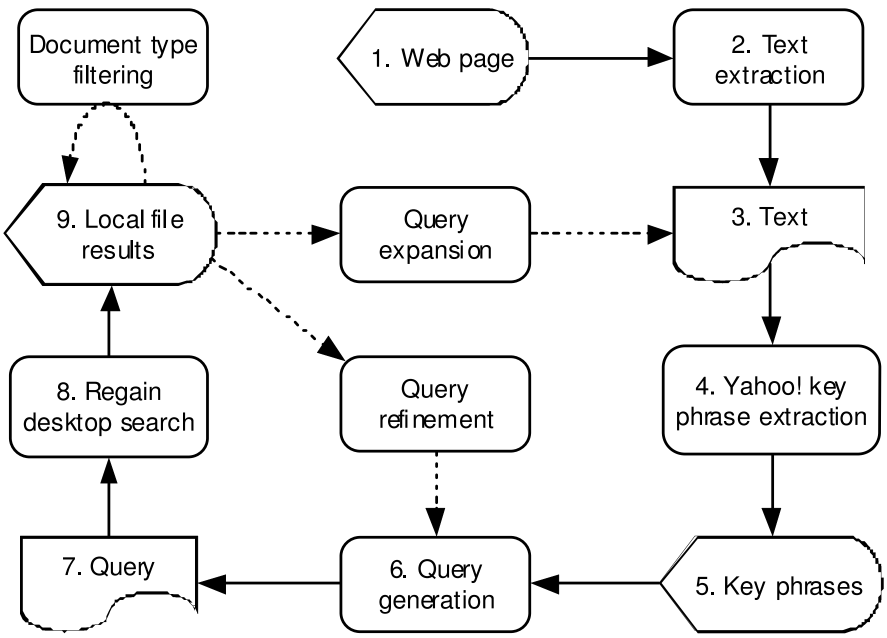


Figure 1 - Context Explorer process

The logging functionality not only gathered information on user behaviour but also logged errors to determine if there was a gross variation in how often the system was functioning correctly. The following errors were all automatically logged:

- Pages that could not be processed owing to a lack of content.
- How often the local indexer failed to respond.
- Https/file protocol connections not processed owing to ethical concerns.

- How often the Yahoo! Term Extractor failed to respond.

5.2 User group profile

The users were profiled in order to better understand the group as a whole and to compare users with a group of 70 college students profiled by researcher Ulla Bunz (2004).

The group for this dissertation was dominated by advanced computer users, with almost two-thirds of respondents indicating they could build a web site well or very well. All respondents had taken at least 1 computer course, with an average of 9.5 computer courses. Almost 90% of respondents used the computer daily at work, and 60% used it daily at home as well.

It was also a very internet-connected user group, with all respondents launching a browser at least once per day and over 70% launching more than 9 times per day. All respondents used the browser for at least 7 hours a week, with almost one-third of respondents using it for over 40 hours a week. All respondents reported that they were "very comfortable" using the Internet.

Comparisons were drawn with the students profiled by researcher Ulla Bunz. Even accounting for the approximately 5-year time difference in the two studies, the user group for this dissertation was a more computer-oriented group. In the Bunz study group, more than half had taken two or fewer computer courses compared to the 9.5 average for the dissertation group. Remarkably, the average hours per week of browser usage were about 10 times greater for the dissertation group (20–40 as opposed to 2–4). This can at least be partially explained by changing patterns in browser usage. The study group for this dissertation also scored much higher in computer comfort levels, saying they were "very comfortable" using computers (70% vs. 45%) and "very comfortable" using the Internet (100% vs. 56%).

A final measure employed by Ulla Bunz was a 12-question cluster designed to specifically measure levels of computer *expertise*. For this cluster, her student group showed 20% of respondents as having expertise/expert status, while the *Context Explorer* user group showed for this cluster a level above 80%.

Overall, the comparison with the Bunz group of college students showed that the study group for this dissertation were using computers for a greater number of purposes and in a more focused manner, were generally more computer proficient and were also using computers for a greater duration each day. This result was not surprising considering that most of the user group worked in information management.

5.3 Empirical investigations

The investigations can be divided into four broad groups. The first considers task context: when is the application useful? The second examines user types: for what users is it most useful? The third looks at result relevance: what factors influence relevance? The fourth and final group looks at various manual intervention strategies to raise relevance.

For a quick summary of results, please refer to *Appendix A – Empirical results summary*.

5.3.1 Task context

The first group of hypotheses examine when and where the application and the scenarios with which it deals can be usefully applied. The first claim was that the "application is more useful for activities that are performed both locally and remotely" – a statement well supported by the evidence:

Criteria: Average number of relevant local files returned by task performed.

- The average number of relevant hits (i.e. the hits rated at least 1% by Lucene) was highest for the remote and local tasks (Figure 2), both work-related. Somewhat lower was the search task (it was noted that many users made work-related searches), and very much at the bottom was the personal task, which had 27% of the relevant hits found compared to the highest-rated task.

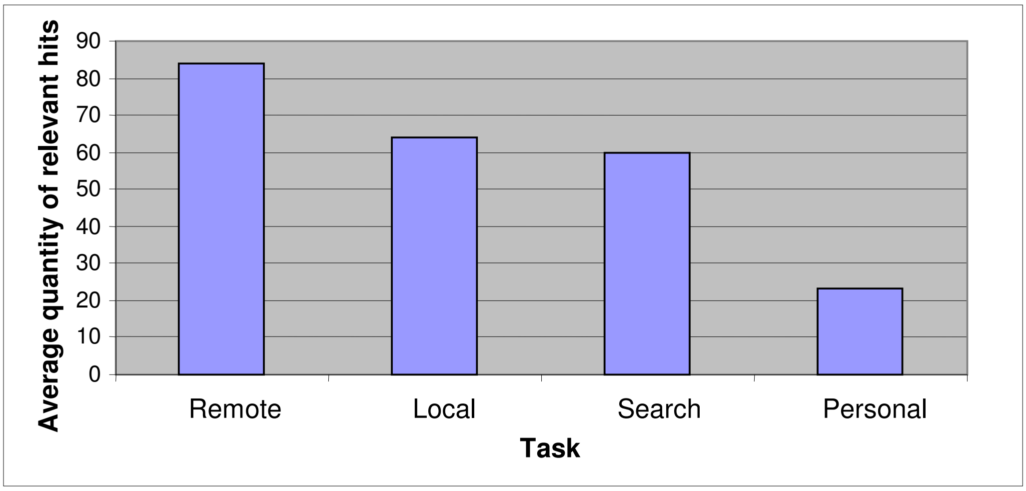


Figure 2 - Average relevant hits by task

- Using a t-test to compare user results one by one instead of as a group showed the variation between the highest- and lowest-rated task to be significant, with a one-tailed t-test probability p value of 0.026 (typically a p value below 0.05 is considered significant). The chi-square test confirmed that the distribution was not random (p = 2.56882E-07).

Criteria: Average relevance rating for returned local files.

- Taking a human-centred approach (as opposed to the machine-generated relevance of the Lucene indexer) created less stark absolute ranges (Figure 3) but extremely predictable user reactions: only 16% rated the personal-task local file results as more relevant than other task results (one-tailed t-tests between highest- and lowest-rated tasks of below 0.005).

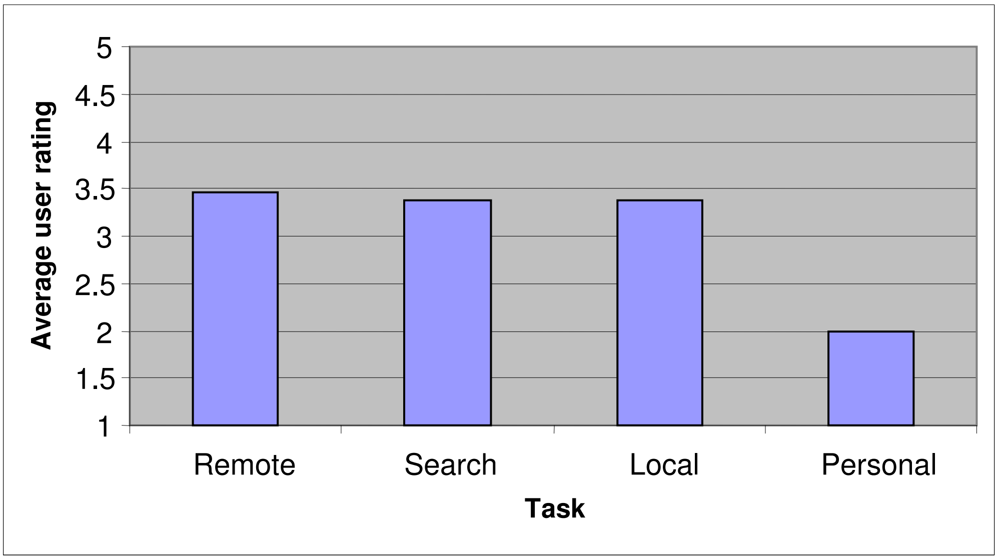


Figure 3 - Average user-reported overall relevance by task

The data clearly supports the argument that tasks have to span local and remote resources in order to give useful, relevant results.

The second context-related claim made was that "fewer results equates with lower use" – a claim that was supported by some of the data, but not strongly so:

<u>**Criteria:** Trend between low hits or small index size and application use.</u>

- The number of page loads by average number of hits was extracted for each user from the logging data for the previous 60 days. No significant trend was found between average hit quantity and number of visits. A similar examination was made using folder quantity to visits, which also did not find a correlation.

- Logging data were extracted (removing test installations and outliers with either a zero index or an index so massive that it made the application unusable) to compare index size versus page loads, which in this case (Figure 4) found a moderate correlation of $R^2 = 0.42$ (towards 0 indicates low correlation, towards 1 indicates high correlation).

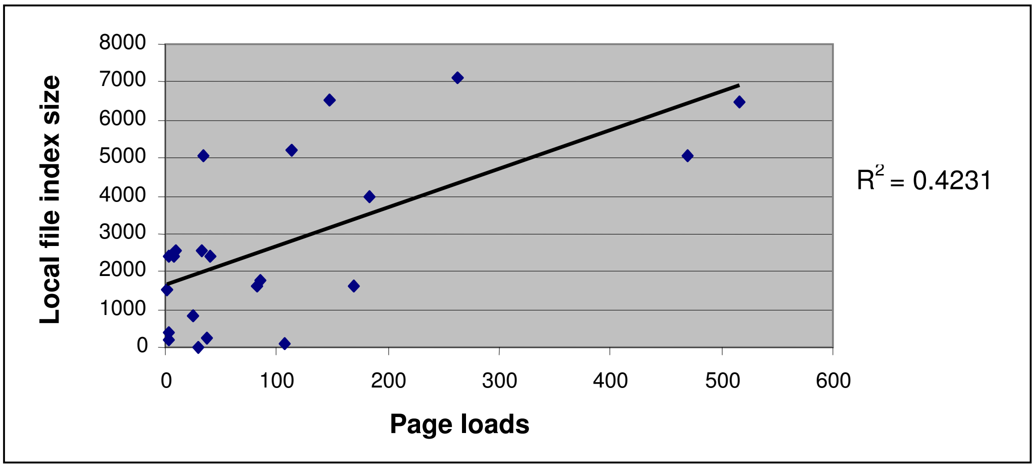


Figure 4 – Page loads to index size

Criteria: Correlation between user rating of relevance and low hits or small indexes.

- An examination of index size to reported relevance during observation sessions showed a similar but weaker trend, with an R^2 of 0.15, while an examination of hit quantity to reported relevance showed no trend (an R^2 of below 0.10).

Overall, a coherent picture emerged between both logging and observation data that, while average hits were not associated with use or relevance, large indexes were. This may be because large indexes indicate a user who has spent more time configuring search directories and so is more apt to use the application.

The final context-related claim was that "there is a correlation between often-visited pages and a greater relevance or quantity of local file results". The data did not support this claim, probably owing to the amount of noise generated by the frequent accessing of search pages, news pages, personal web mail and AJAX desktops:

Criteria: Common pages return larger and more relevant local file sets.

- Using the logging data, over 10,000 URLs were analysed, from which it was determined that there was no correlation between local hits and frequent page loads (Figure 5).

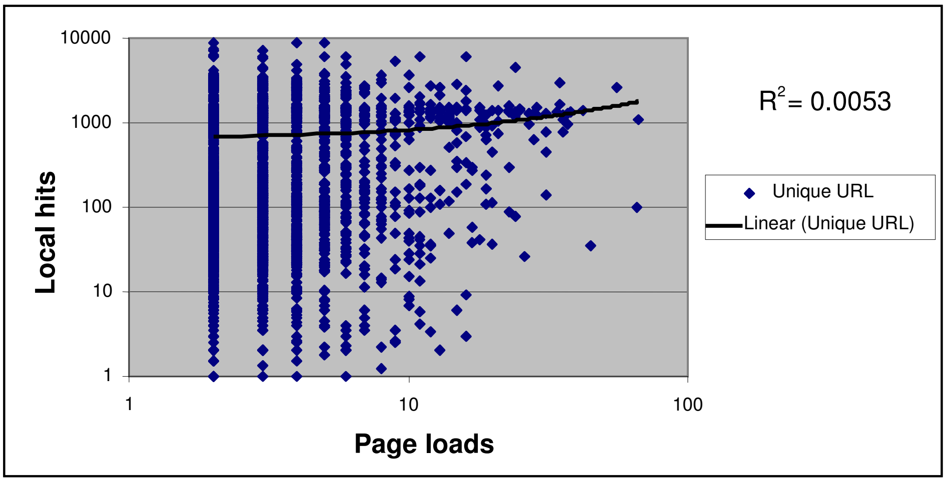


Figure 5 - Page loads to local hits

- Lucene-rated relevance for URLs was also examined against hits received by a URL to see whether higher relevance scores and often-visited pages were correlated, but no correlation was found (Figure 6).

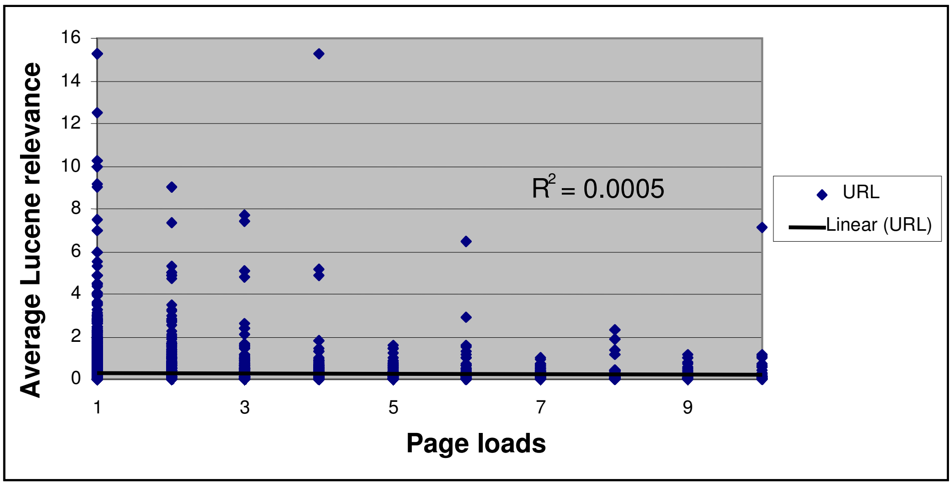


Figure 6 – Page loads by Lucene relevance

The view that noise prevented an analysis of this claim seems borne out by the fact that the number-one URL was *google.com*, with almost a quarter of all hits. In addition, it was difficult to define what a unique page was, as pages change rapidly in today's dynamic, advertising-loaded web environment. An analysis of logging data showed that users visited a page an average of only 1.4 times before its string length changed, indicating that some kind of content change had taken place. An extended analysis of the URLs would be needed to attempt to filter out such URLs.

5.3.2 User type

Three of the hypotheses dealt specifically with user types: What kind of user might benefit most from an application that associates remote and local resources? The first claim was that "users who spend more time organising their data in many folders have more relevant results" – a claim that was not supported by the data.

Criteria: Greater application use by users with a greater number of folders returned.

- There was no relationship between the use of the application and users with a high folder/file ratio (Figure 7), as the R^2 value that shows the strength of the correlation was extremely low, at 0.004.

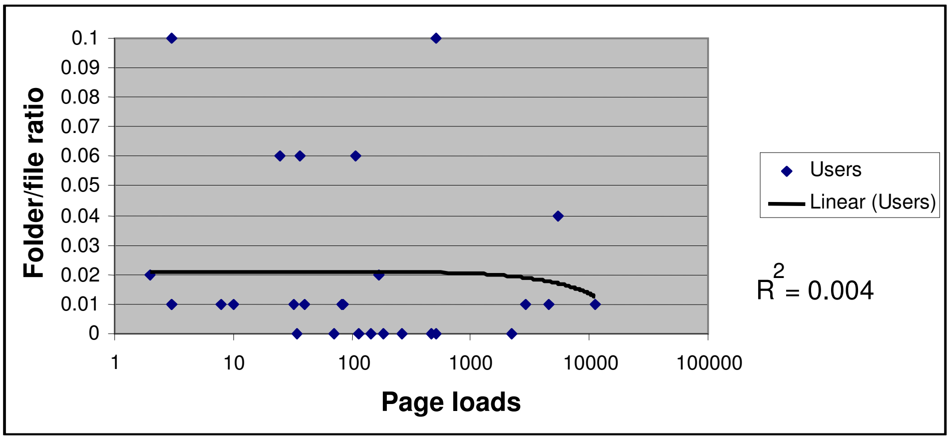


Figure 7 – Page loads by folder/file ratio

Criteria: Trend of higher relevance indications by users with a higher folder/file ratio.

- A series of six distinct relevance types were analysed: between files and activities, files and web pages, key phrases and activities, key phrases and files, key phrases and web pages and overall relevance for the browser interface. These user-reported relevance scores were compared to user folder/file ratios, and again, no correlation was found for any of the relevance types, with a maximum R^2 value of 0.07, which is below the 0.1 that might show even a minimum correlation.

The second user type claim had to do with competing visions of resource: by folder or by file? The hypothesis was that, "for some users, folders play an important role in interpreting results" – a view weakly supported by the evidence for very small and very large results sets.

Criteria: Folder view selection frequency.

- Analysing the logging data, it was found that the folder view was toggled 210 times. Considering that page loads totalled 36,467, folder-view toggling represented only 0.6% of page loads, and the default file view was overwhelmingly preferred.

Criteria: Number of hits returned when folder view toggled versus average number of hits returned.

- The result-set size was on average either smaller (below first quartile) or larger (above third quartile) when the folder view was toggled (two-tailed t-test $p = 0.042$) – a finding also supported by observation sessions.

Criteria: Users indicate they prefer the folder view.

- When users were asked, they overwhelmingly preferred the file view, rating the folder view as superior only 25% of the time. The data were analysed using least-squares to determine whether user choice bore any relation to the index size, the number of returned hits or the number of Lucene-relevant hits. Stratifying the results, there was found to be a low correlation to small result sets, both below the median ($R^2 = 0.11$) and above the third quartile ($R^2 = 0.12$) – a finding that matched the logging data results.

The third user type claim was metadata-based: How much information do users need in order to recognise their file results? The claim was that "users recognise files by their filenames", which turned out to be overwhelmingly true.

Criteria: Is metadata necessary in order to recognise file results?

- Users strongly indicated (4.07 out of 5) that they recognised their files by filename. This was true even for users with 10,000+ indexed files (Figure 8). A trend analysis showed a strongly significant trend ($R^2 = 0.87$) in user responses.

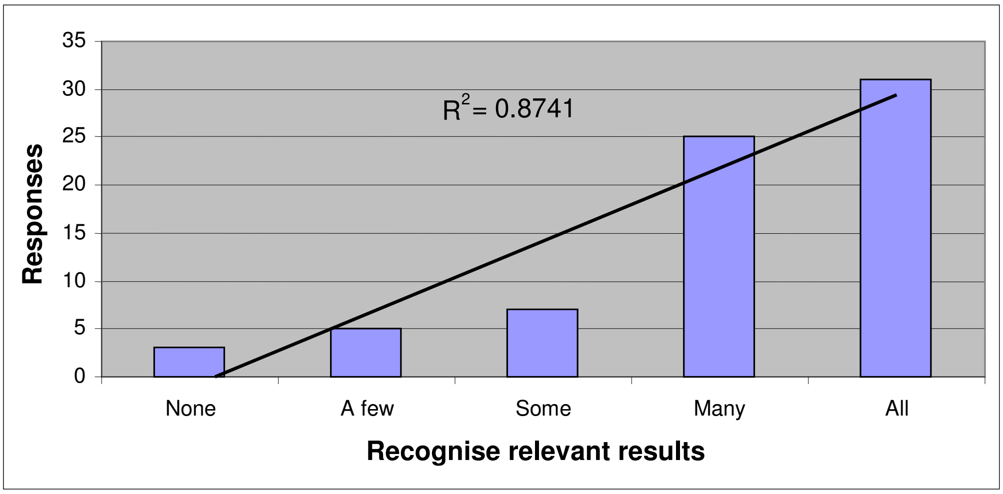


Figure 8 – User recognises files

- The more files a user had indexed, the more they confidently claimed they recognised all files by filenames (Figure 9). The correlation was moderate, with an R^2 of 0.37.

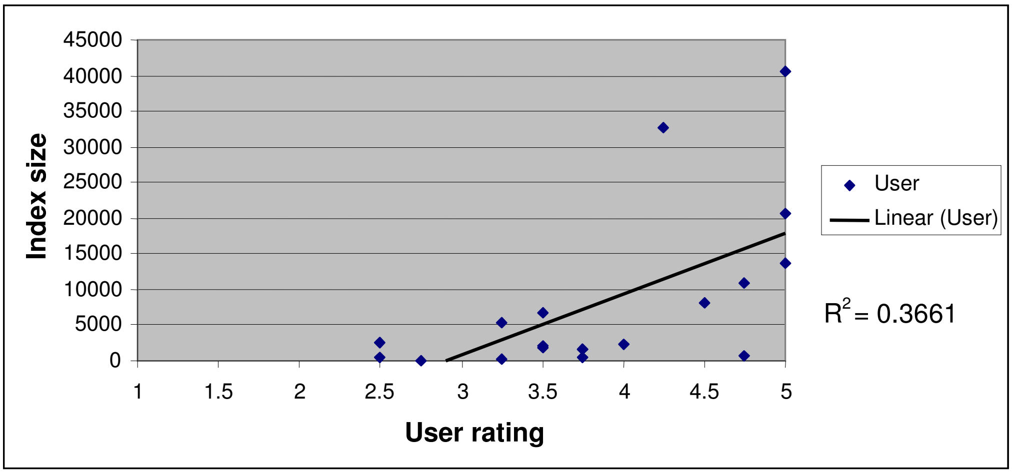


Figure 9 - Recognise files vs. index size

- Users also found that, although they did not generally need it, a summary consisting of the first 30 words in a file was very helpful in identifying the file, with users rating the utility's helpfulness as "very" or "excellent" 87% of the time (Figure 10).

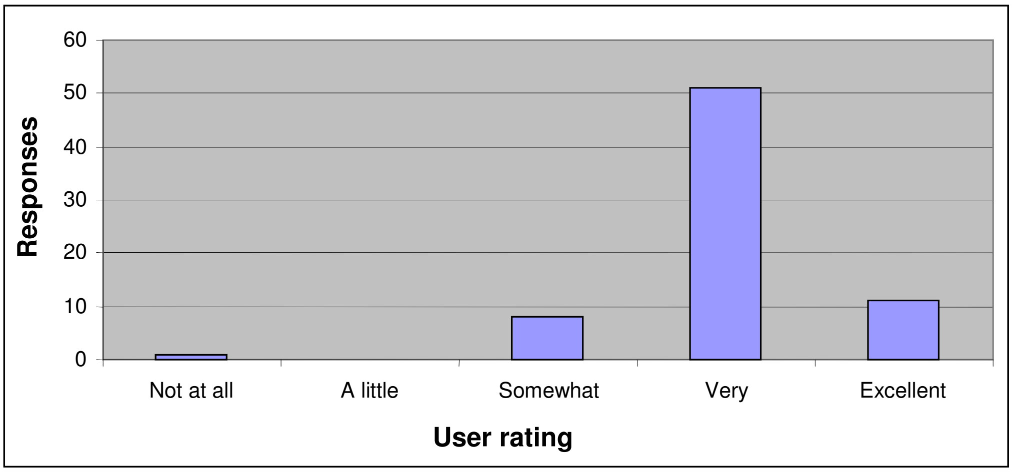


Figure 10 - Utility of summary metadata

In light of the user appreciation of summary data, it is interesting to note that *query expansion* – a technique that uses file summaries as a way of extracting key phrases to generate more local file results – was rated quite negatively.

5.3.3 Relevance factors

The third set of hypotheses dealt with factors inherent in the data that might affect relevance. The first claim was that "only the most relevant results are at all relevant". It was felt that the combination of the broad scope of some web pages and the imperfect method of key-phrase extraction would create a situation in which only the most relevant results would be relevant – a hypothesis that was well supported by the evidence.

Criteria: Relative frequency of the use of the "next 50" results button.

- The average number of hits when the next/previous features were used was very high – 2,752 – compared with an average of 523 hits (a one-tailed t-test gave a value much lower than alpha – 1.66E-07), which shows that users were more likely to peruse the next 50 results when they knew that many hits were returned. However, the feature was seldom used and accounted for no more than 0.3% of all page loads.

Criteria: Higher average relevance when the system was used in any way.

- There was a marked difference in average Lucene relevance when the system was used in some way (0.472), compared to the average relevance for all page loads (0.235). A chi-square test confirmed that, for the 1,243 logged events, the distribution of relevance was not random. Clearly, users were using the system more often when the relevance of the results as rated by Lucene was higher.

Criteria: Higher machine-rated results were scored more relevant by users.

- In terms of Lucene-rated relevance (based on the classic vector-space model) and whether users agreed with it during observation sessions, they generally did. When asked if the "black file results were better than the grey file results" (the application coloured grey any result that had lower than 1% relevance), users tended towards rating it as "better" or "much better" 54% of the time, "same" 37% of the time and "worse" or "much worse" only 9% of the time (Figure 11).

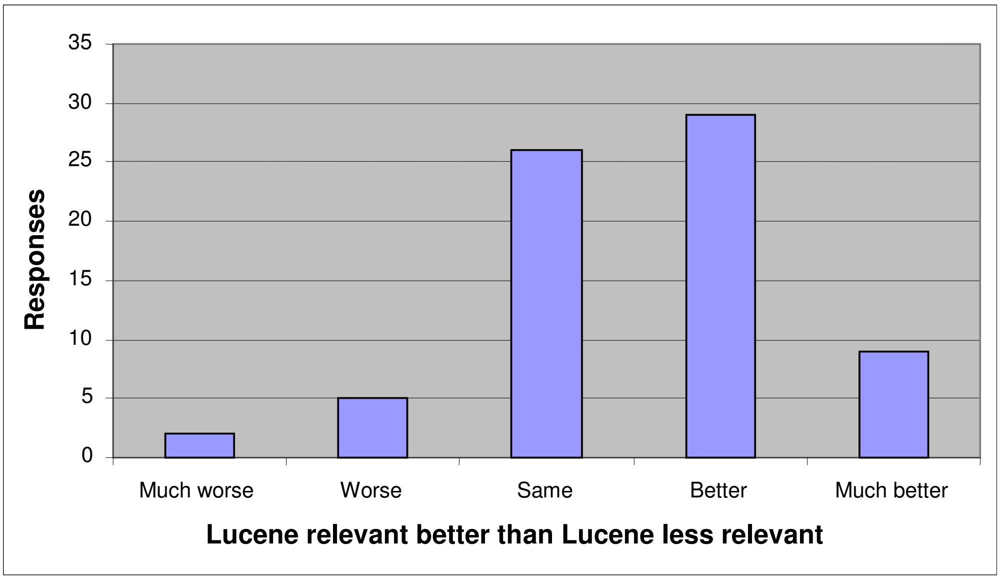


Figure 11 – User-scoring of Lucene vector-space model relevance

- Interestingly, this was with reference to the relevance between *local file results and web page*, not between *local file results and key phrases*, which shows that, despite the several steps of processing (web page → key phrases → query → local file results; see Figure 1 – Context Explorer process), users still felt that relevance had not been lost. That it was diluted is clear; users typically scored the relevance between the 2-steps removed *local file results and web page* lower than the 1-step removed *web page to key phrases* or *local file results and key phrases*. Nevertheless, the responses to a simple binary question of whether the higher-scored or the lower-scored local results were more relevant to the web page were overwhelmingly positive.

The second claim made for inherent relevance had to do with the scope of the web page in terms of its length or key-phrase characterisation. The hypothesis was, "the more key phrases extracted from a page, the lesser the relevance of the results", for which weak support was found.

Criteria: Trend between key phrases or page length and lower average relevance.

- The pages were aggregated by the number of key phrases they generated (stopping at 19, as the key-phrase generator stopped at 20 for any page, skewing any results above 19), and the average relevance was calculated for some 20,000+ page loads. A moderate correlation was found (R^2 = 0.33), with strong average relevance for low quantities of key phrases, rapidly tailing off past five key phrases (Figure 12). This result is due to the higher probability that any document could match a greater percentage of key phrases when there are few of them.

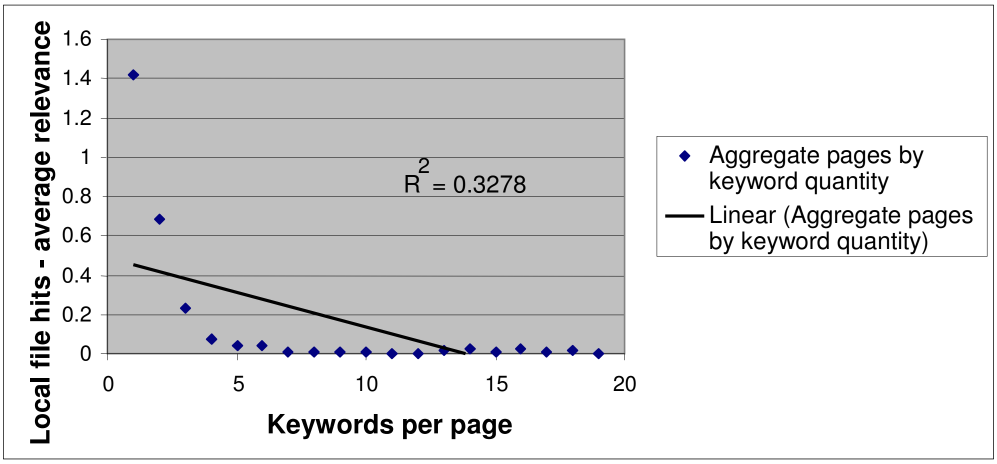


Figure 12 - Lucene relevance by key phrases per page

- A similar analysis was made examining Lucene relevance to page length. The assumption was that longer pages were less focused and would generate lower average relevance for local file hits. However, no relationship was found (Figure 13).

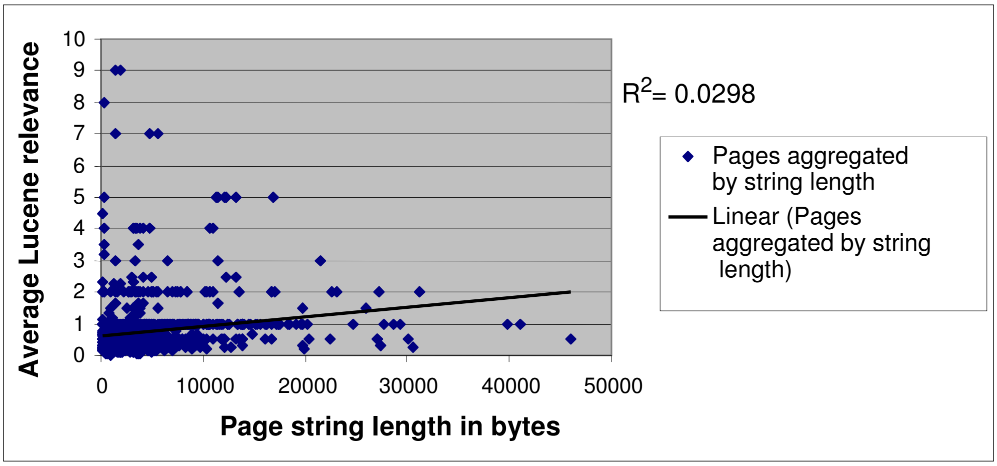


Figure 13 - Lucene relevance by page length

- Further analysis clarified the probable reason for this. When page length versus key-phrase quantity was also analysed (Figure 14), it was found that there was no relation between the length of a page and the quantity of key phrases it generated.

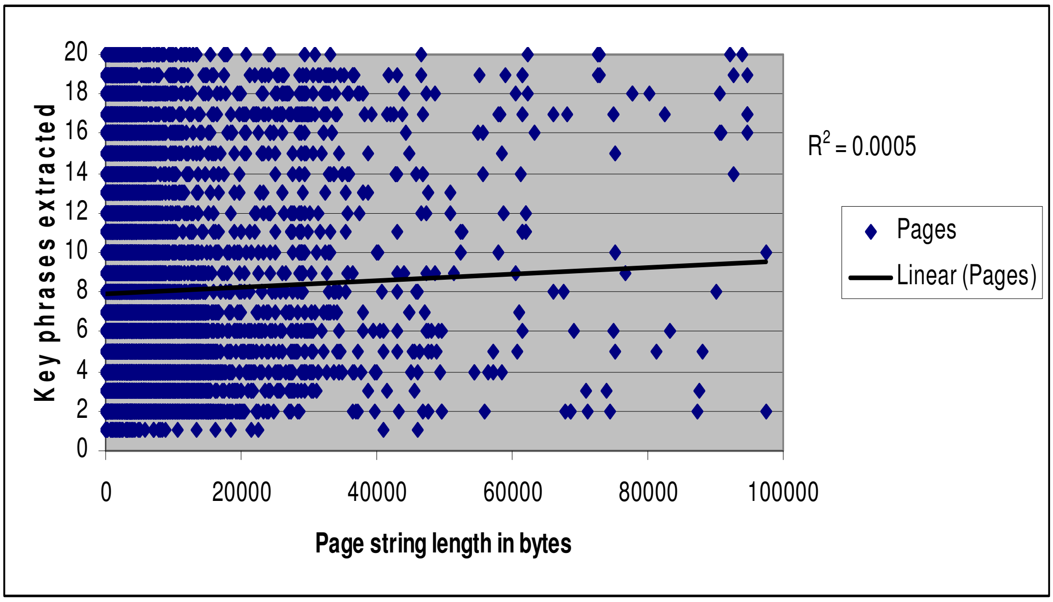


Figure 14 - Number of key phrases extracted by page length

<u>**Criteria:** User relevance rating compared with key-phrase quantity and page length.</u>

- User observation did not find the moderate correlation indicated by the logging data, a discrepancy due to small key phrase sets better matching documents as noted in the previous criteria. User observation relevance scores for the overall relevance of a page load were compared to both page length and key-phrase quantity. No relationship was found, even when comparing low numbers of key phrases to high numbers of key phrases.

The third case made for inherent relevance had to do with key phrases, specifically that "frequently occurring key phrases are associated to more relevant result sets" – a claim that was totally unsupported by the evidence.

Criteria: Trend between common key phrases and actual system use.

- The 15,000+ unique key phrases extracted by the logging system were analysed against average relevance data for the pages on which the phrases were found. No correlation was found between commonly extracted key phrases and average result-set relevance (Figure 15).

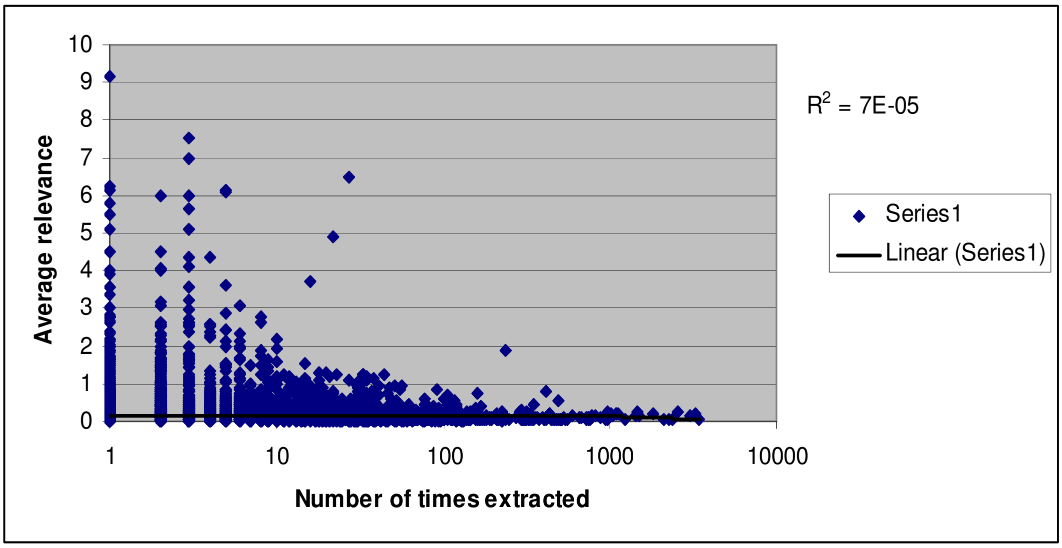


Figure 15- Key phrase frequency to average relevance

- A cursory examination of common key phrases showed that many were simply products of web search interfaces or web mail programs, e.g. "Google", "Lycos" or "Gmail". It is thus quite understandable that no relationship was found. The application of a set of stop-words to filter out such key phrases might change these results.

5.3.4 Manual intervention

The last group of claims centred on the possibilities of how manual intervention could improve the relevance of the system. The first claim was that "filtering is a useful selection mechanism to increase relevance", which turned out to be clearly the case.

Criteria: Correlation between result-set size and the use of filtering functions.

- At first glance it seemed tempting to associate the use of key phrase and document-type filtering functions to large result sets, as the mean for hits returned when filtering functions were used was triple the mean for all hits returned. However, a t-test analysis showed an alpha of around 0.15, so that, although some effect was found, it could not be considered significant, as it was above the threshold of 0.05.

Criteria: Higher or lower average relevance ratings for filtered documents compared to relevance when viewing all documents.

- Figure 16 shows a strong preference for the relevance of Microsoft Word documents (.doc and .rtf extensions) compared to the average value of 3.10 when document-type filtering was not applied, while all other files scored below average.

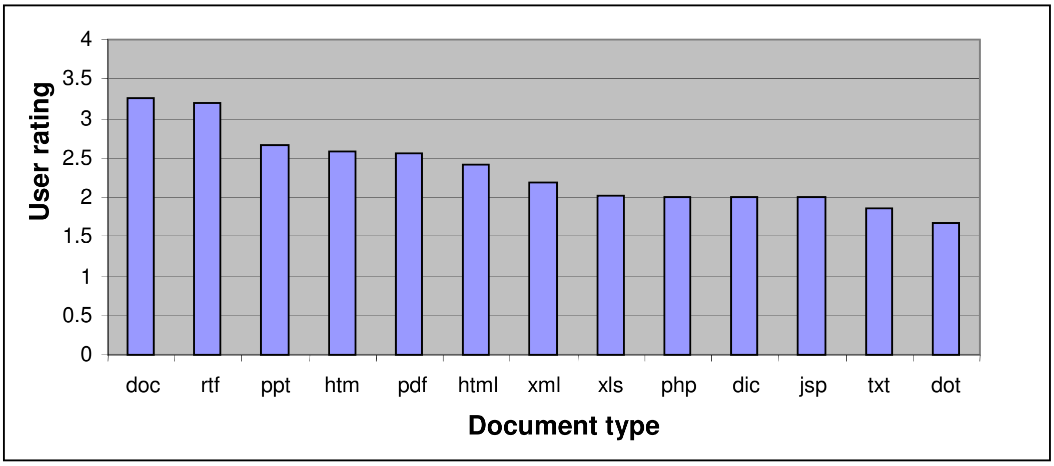


Figure 16 - Average relevance by document type to web page

- A chi-square test of the 241 results from the user observations confirmed that the variance was not random, and a t-test analysis comparing relevance ratings between filtered and unfiltered views gave a two-tailed *p* value of well under alpha at 0.002, confirming that the difference in relevance was statistically significant. Filtering by document type was important to the users' indication of relevance, though it does not appear to be linked to a function of result-set size.

The second claim made for manual intervention had to do with content extraction selection. The hypothesis was that "key phrases extracted from title and metadata create more relevant results" – a statement that produced conflicting data between logging and user observation.

<u>**Criteria:** How often users employed content, title or metadata to generate key phrases and the local-file result relevance generated.</u>

- The logging data showed a very strong preference for the "All" method, although this method showed the lowest average relevance (Figure 17), probably because users change extraction type only when they are getting to something interesting – otherwise it was left on the default "All". The higher relevance of the "Title" method is almost certainly due to its brevity, causing it to generate few key phrases.

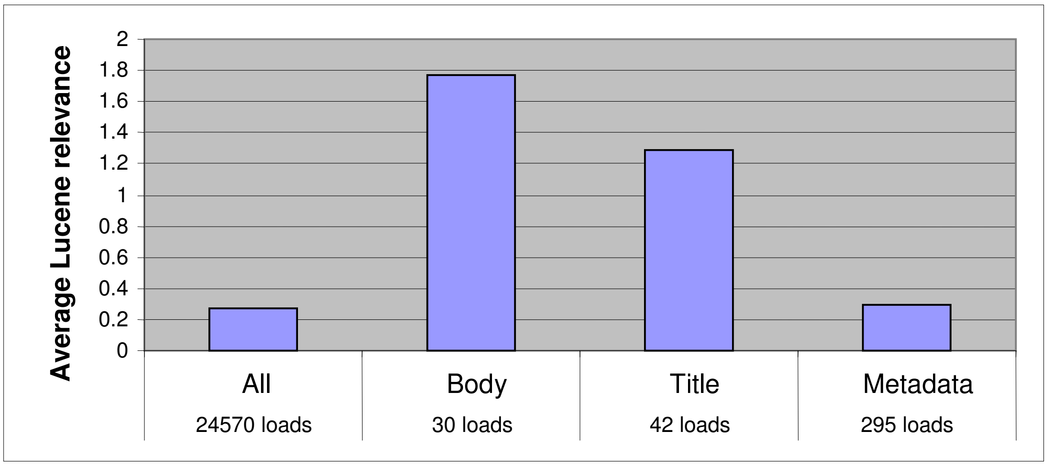


Figure 17 - Content extraction methods

<u>**Criteria:** User indications of relevance during observation sessions when asked to switch between content, title or metadata.</u>

- Users were mostly indifferent to the various extraction types when asked whether the relevance of the file results to the web page had improved when trying a new content extraction method: *title* extraction tended slightly towards "worse" and *metadata* extraction slightly towards "better" (Figure 18). This is probably owing to the extremely limited semantics that a title provides for a web page.

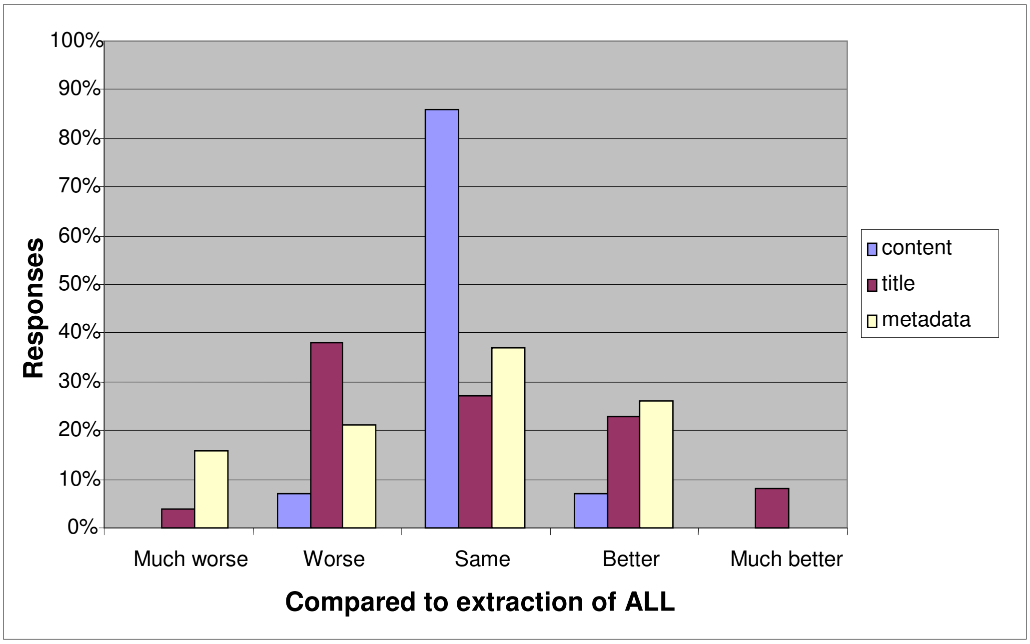


Figure 18 - Relevance by extraction type

Lastly, a form of automatic relevance feedback was examined, the claim being that "query expansion using automatic relevance feedback is not very useful" – a statement the data was very much in agreement with.

<u>**Criteria:** How often users employ query expansion.</u>

- Query expansion was less than 1% of page loads; it seems not often used and so probably not very useful.

<u>**Criteria:** User indications of query expansion relevance.</u>

- The observation sessions provided more direct feedback, as users were asked to perform query expansion in two distinct scenarios for each task: one where the expansion was based on the top-rated documents as a function of all the key phrases extracted, while the other asked users to perform a query expansion after they had narrowed the result set to a single key phrase. These two types of query expansion performed similarly. Respondents rated the result of query expansion as worse or much worse 45% of the time and better or much better only 21% of the time, while query expansion based on a single preferred key phrase helped a little, lowering the worse/much worse score to 34% and the better/much better to 27%. This is still a fairly poor showing, however.

- The results using the best key phrase were much more encouraging, with respondents rating the results as worse only 15% of the time and better or much better 52% of the time, showing that characterising a web page by a single, user-selected key phrase generally led to higher relevance in the file results.

5.4 Validation

The research question regarding the practicality and utility of the automatic association and filtering of local resources to web pages was at least preliminarily answered by the research. Hypothesis testing was generally successful, and there were few cases that could not be either shown to be statistically supported or unsupported. The case study has shown the technique is practical, and for certain tasks the association between local and remote resources is quite relevant.

A lingering doubt nevertheless remains as to how *useful* is the technique. Users indicated strongly that the relevance for work and research tasks was high, but this does not necessarily translate to the technique being useful. The hypotheses that were tested showed that the technique was successful in associating local and remote resources, but to ultimately test if this is *useful* would require additional time. The application was released to shareware sites in autumn 2007, and logging data shows the application in use by dozens of users. This data would need to be monitored over a longer period of time to see if the user base continues to grow to ultimately determine the real-world utility of the specific approach developed.

5.4.1 Method analysis

The quantitative logging and qualitative observation methods were appropriate to the situation, though some limitations applied. Direct user observation using talk-aloud protocols suffers from the possibility of influencing user behaviour by the participation of the experimenter in the process (Open University, 2001). The logging data did not suffer from this defect, which may have caused problems when comparing results from the two methods.

Though the primary data collection based on indirect (logging) and direct (task scripts) user observation with its mix of quantitative and qualitative data collection was successful in providing the necessary data to test the hypotheses, secondary data was notably missing. There were several reasons for this. As mentioned in *3.3.1 Primary and secondary data*, the innovative application design meant that other examples were not found for comparison. At a lower level, given a longer study, it would be possible to compare primary and secondary data such as average query relevance or precision between the application's automatic queries and other manual query scenarios. The same could be done for query recall data, but it was sensibly decided early on that these comparisons were too ambitious for the scope of the study.

Other methods could have been more amenable to comparison with secondary data. Instead of creating a new kind of test-bed application that was unable to be directly compared to secondary data for other applications, questionnaires could have been designed to study user attitudes towards the link between local and remote resources. These could have been compared to existing studies. The strengths and weaknesses of existing search strategies and applications could have been analysed and compared to a proposed self-adapting local/remote explorer.

Another method that might have been just as useful as creating an application would have been to test a theoretical application using low-fidelity paper prototypes. The advantage here would have been the possibility of creating best-possible-outcome scenarios that would not have suffered from any real technical limitations. This could have created a clearer evaluation of the concept by better separating the theoretical aspects from their practical implementation.

5.4.2 Design analysis

Data collection design consisted of three main areas: logging data, user observation and questionnaires. While the user observation and the survey design were carefully plotted prior to implementation, this was not true for the logging design.

The test application was designed and implemented early in the project to insure that this critical component could actually be realised. Given its tight coupling to the centralised database, the logging data design was completed without extensive testing to see how difficult the data extraction for hypothesis support might be. This created some difficulty in the later analysis phase, and some data that could have provided additional hypothesis support proved too hard to extract, while other unneeded data were also collected. A further weakness was that, owing to privacy concerns, logging data and user observation data were not directly compared for the same users but for the group as a whole.

The user observation sessions provided the most focused data for answering the hypotheses and were vital to meeting the project's aim. In contrast to the logging data, these were piloted through several rounds before being finalised. This increased the correspondence to the hypotheses, and data were easy to extract for analysis.

The survey design was quite adequate in itself, but due consideration was not given to how the survey data could be analysed with logging and observation data to draw conclusions about user types. In practice, the survey data added little or nothing to hypothesis testing and served only to profile the group as a whole.

5.4.3 Statistical power

While the logging data were drawn from a larger pool of users (over 30), the sample size for the observation sessions was 19 users and for the profiling survey 14, which was not enough to be considered statistically relevant. However, an analysis of t-test degrees of freedom (a measure of statistical power) showed that the difference between the user group of 19 and the desired group of 30 was minimal. Nevertheless, the study results must be considered to be preliminary.

Another limitation was that the group was not a random group but was selected for convenience from amongst friends and co-workers. This group suffered from several problems pointed out by the user profiling, in particular that it was overly represented by very experienced computer users.

6 Conclusions

This project looked at mixing aspects of computer- and human-centred approaches to see how they could bring users' fragmented vision of their local and remote computing resources closer together. Although further research with greater statistical power is needed, the research has preliminarily shown the practicality and utility for some tasks using a self-adapting interface that dynamically associates local resources with browsing activity.

One of the central objectives was to examine the utility of such tools based on the situation or task in which the user was engaged – an approach favoured in HCI research on minimalism (Hackos, 1999). It was felt that the utility would be strongly dependent on the type of activity. This claim was well supported by the research. Work and research tasks related to activities that spanned users' local and remote resources generated much greater relevance both with human- and with computer-centred measures, while for personal tasks – surfing news sites, reading blogs, etc. – the tool was not only considered useless but also annoying, as it took up screen space and added a very slight drag to the speed of web page loading. The implication was that the tool was only useful some of the time, and that such an interface needs to be easily dockable when not needed, perhaps self-docking when it has not been used for a certain period.

Within the assigned tasks, another objective was to see how interest in the tool differed based on the users' differing computing profiles. Much recent usability research has pointed to the differing needs of various user groups, from newbies to the differently-abled. User observation sessions showed that the tool was rated more highly by power users who had large amounts of local files. They were also more likely to use the plug-in. Somewhat paradoxically, it was also these power users with enormous indexes of files who were generally more confident about the identity of the result sets that appeared in the application, perhaps because large indexes force a user to spend more time organising files and adopting naming conventions. The implication was that users with small indexes might benefit less from such a tool and that when using it they may require more hints about file identity to understand the result set, such as a hierarchical file/folder view, which some users requested during observation sessions.

The role that folder hierarchies play in aiding users to understand their resources was another area of investigation, given the current buzz surrounding semantic data organisation that claims to leave hierarchies behind (Xiao & Cruz, 2005). It was thought that result sets organised by folder might be the preferred method for many users. Research, however, showed that not a single user showed a marked preference for folder views, and such views were deemed more relevant only about 25% of the time, generally when result sets were poor and the folder view acted as a way of expanding the result set, or when the result sets were very large and folder views acted as a grouping mechanism. The implication is that folders and hierarchies are not as important as one might think and that the trend towards semantic and relational file systems is probably correct. Classic hierarchic folder visualisation could probably be left out of file-exploring tools given good queries and rich semantics, though folder text might be used in the background as a relevance boosting mechanism where it matches query key phrases.

A touchstone of the research was interface self-adaptation: the idea that the user is the environment and that, like chameleons, systems should adapt themselves to that user environment, shaping themselves to users' needs. One of the most evident conclusions drawn from the research was that such approaches are not yet mature. Some amount of user interaction is often necessary to raise relevance to an acceptable level. Some of the most compelling data produced was related to user-initiated filtering. Relevance scores were unquestionably higher when users were allowed to filter by document type or by key phrase. Interestingly, users more often than not scored local file results as being more relevant to a web page when they were allowed to select a single key phrase from the key phrases extracted for that web page. The implication was that the single most relevant key phrase for a web page often led to the most relevant set of local files – a surprising result that requires further research for us to understand how this characteristic can be better exploited.

A computer-centred IR technique for raising relevance without manual intervention was attempted, but whereas manual query refinement was very effective, the research performed on automatic query expansion performed dismally. This was not surprising however, as this type of automatic relevance feedback has a long history of mixed results (Baeza-Yates & Ribeiro-Neto, 1999). The application used a variant of the method, which took the summaries of the top five ranked results and used them to regenerate a set of key phrases that then re-queried the user's local file index. While it occasionally performed brilliantly, raising the relevance of the result set for the user, in most cases it diluted it, even when it was based on the documents returned from the user's single, favourite key phrase. This again reinforces the previous implication that some amount of user interaction is often necessary to correctly contextualise resources.

The most semantically correct method to characterise a resource has long been a point of contention among IR experts. Approaches vary from human indexing with metadata to vector-space modelling to newer techniques such as LSI preferred by semantic desktop supporters (Dumais, 1988). Though exotic techniques such as LSI were not attempted here, the project tested various web page content extraction techniques for differences in relevance by using just the web page title, just the explicit header metadata (when available), just the content or all of the above.

These content extraction-type comparisons were the only research area in which relevance differed markedly when comparing logging and observation session data. In analysing the logging data, as far as machine-relevance was concerned, results were best when using the web page title – an effect caused by the restricted number of key phrases generated by titles compared to the number of key phrases generated by an entire page of content. This facilitated a higher average relevance since it was more likely that any one document would contain the few key phrases generated.

Users, however, were less impressed with extraction based on page titles, rating this method lowest of all. This outcome was not surprising. What it showed is that, while the web page title generated a set of local file results that were quite relevant to the web page title, they were not as relevant to the web page content. The machine relevance had no way of measuring this misfit between content and title, whereas users implicitly did. There is room here for future work, e.g. double queries could be made and relevance assigned based on a relevance mixture model between methods. One can infer from this outcome that web search engines are correct in their general disdain for web page header metadata such as titles, often treating it no differently from the rest of the web page content. However, this reliance on visible page content means that it is even more vital to strip advertising and menu structure from actual page content to improve key phrase extraction.

Although the machine relevance was inaccurate for title-based queries, this was an isolated case. In most cases the average relevance in the result set was a vital marker of whether the system was actually used. It seems that only the most machine-relevant sets were relevant at all. As part of the objective of comparing the computer and human ratings of relevance, the logging data of the average relevance when the system was interacted with was extracted and found to be more than double the average relevance for all ignored page loads. This fact accorded with the observation session result in which users overwhelmingly agreed with the simple binary question of whether the higher machine-ranked results were better than the lower ranked results. The implication is that, at least at a level of low granularity, users very much agree with the relevance returned by the vector-space model – which is a reassuring result, as this model forms the heart of most successful search engines.

A final objective of the research was to examine user browsing behaviour for very commonly accessed pages and whether there was a strong relation to local resources. No relationship was found. The assumption was that pages often visited by users would be naturally related to common activities that would be in turn related to their local file resources. However, probably because the most common pages were extremely generic, such as search pages, web mail pages and news pages, no effect was found. This same non-effect was found in analysing commonly occurring key phrases, e.g. "mail", "web" and "Gmail", which turned out to be useless in generating relevant local results. The implication is that the most common pages in fact have little to do with user resources – they are merely generic gateways. An interesting future development would be to build a set of stop-URLs and stop-phrases to squelch noise and to determine whether there is in fact a more powerful relation to be found between second-tier pages or key phrases and local file results, i.e. pages that are visited often but not always.

6.1 Project review

The overall project aim was quite ambitious for the one-year time frame in which it had to be carried out. Effectively assessing inherently fuzzy terms such as "practicality" and "utility" meant an elaborate research setting, a fact that was not sufficiently considered at the beginning of the project. In order to test usability aspects, the plug-in created for testing had to be more mature than would typically be expected for a research project. This meant significant up-front development effort, and without the availability of several powerful existing components that were cobbled together to reduce development time, the entire project might have failed.

6.1.1 Quality control

A subset (20%) of user observation logbooks were double-entered to check for transcription errors. Perhaps owing to the database and input form constraints that were put in place, no transcription errors were found.

As logging data were streamed automatically from application instances to the central logging database, transcription errors were highly unlikely. During application testing, user actions were compared against the logging data to ensure the system was recording data correctly. Error conditions were classified and recorded. The percentage of pages that could not be processed was 19%. This figure includes the 6% of pages that were purposely not processed for ethical reasons (HTTPS and FILE protocols) and the 7% of pages lacking enough content to be processed (login screens or pop-ups), leaving a real failure rate of 6% caused by either the key phrase extractor (2%) or the local indexer (4%).

6.1.2 Technical limitations

Some technical limitations were found during collection and analysis. The first was the problem of extracting content from a web page that a normal user would consider to be actual content. The extractor could not distinguish between normal text and advertising or menus: key phrases were extracted from both, diluting key-phrase relevance. Secondly, modern web technology makes it increasingly difficult to recognise page loads. The use of AJAX techniques that update parts of a web page without reloading made it difficult to catch all loads.

Though not strictly a technical problem, for ethical reasons the 6% of page loads consisting of HTTPS and local FILE protocols were not processed. Though some page loads were deliberately lost this way, there is no particular reason to suspect that user behaviour differed for these network protocols.

As discussed in *3.1 Research setting*, a limitation of the Yahoo! key-phrase extraction was that the algorithm used is commercial property and unpublished. In addition, extraction is limited to a total of 20 key phrases, meaning that, for very long or semantically complex pages, some meaning was not extracted. In fact, some users were surprised to note the absence of key phrases that they felt certain should have been extracted from a page.

6.2 Future research

Some interesting avenues were not explored, and some problems in the execution of the methods were encountered that could be resolved by subsequent research. Opportunities fall into several areas: web page content extraction and characterisation, local-document indexing and query results, the inter-session monitoring of user behaviour and enhanced application features.

As seen in *6.1.2 Technical limitations*, web page content extraction proved imperfect owing to the advertising present on many pages. Techniques to filter such advertising from the characterisation of a page by using stop-words would be an interesting area of research, as would acting upon formatting information in the page, such as bold text and large fonts, to emphasise certain key phrases. In addition, the suppression of some common URLs such as search pages and web mail clients would reduce noise in the data and refine results. Particularly interesting would be the possibility of allowing users to build such stop-lists, which could also then be shared to create a folksonomy of URLs and key phrases generally irrelevant to local resources. This could then be tested to see how and whether it raises relevancy for new users working straight out of the box.

Characterising a page by extracting key phrases is a process in which some semantics are lost. Future research could test other methods either alone or in combination with the already tested technique of key-phrase extraction. The use of social bookmarking tools such as del.icio.us, in which users tag pages with key phrases, is adding an entirely new layer of meaning to web sites and web pages. These key-phrase folksonomies could be combined or used in place of key-phrase extraction.

It would be useful to experiment with and compare other term extractors. An open-source term extractor with a public algorithm would permit greater key-phrase analysis.

In addition, the corpus training features of many extractors could customise extraction to the individual user instead of returning a generic document characterisation – a missing feature that was both a strength and a weakness of the Yahoo! Term Extractor.

Local document indexing and querying is another area in which the research could extend. In the current project, the vector-space model was used. It would be interesting to experiment with LSI, moving the key phrases and the local resources into a shared conceptual space. Folder hierarchies might also be exploited. The names of the folder and parent folders in which a resource is found can function as an additional resource in assigning relevance to results, as can the relative weighting of results by mixing the sets arising from title, metadata and content queries. This approach might also be tested to provide a more balanced result set based on a mix of web page components.

Inter-session behaviour is an extremely interesting angle that was not touched upon in this project. In addition to the previously mentioned stop-phrase and stop-URL lists, tracking user session behaviour and modifying results based on past behaviour is a strategy that could prove fruitful. Using techniques of boosting for both key phrases and local results when selected could act to increase relevance over time. Boosting key phrases that are extracted often for a user could also be interesting. This would permit data analysis that with use could create a truly task-aware application, capable of guessing when it is not needed.

In terms of the plug-in itself, there were a number of features that users requested and that could add interesting research angles: additional filtering by date, alphabetical sorting and views that combine folder/file hierarchies. Some users requested a query box within the plug-in that would allow them to actively query within a result set. These terms could be used to build a custom, local folksonomy associated with key phrases and key-phrase sets.

Finally, as mentioned in *6.1 Project Review*, user profiling from surveys was not at all linked to logging or user observation sessions. This analysis could provide a more accurate view of how user experience and user patterns affect the utility of contextual interfaces that relate remote and local resources.

References

Aula A. and Käki M. (2005) 'Less is more in Web search interfaces for older adults', First Monday, vol. 10, no. 7 (July) [online], http://www.firstmonday.org/issues/issue10_7/aula, [Accessed 20 March 2007].

Baeza-Yates, R. and Ribeiro-Neto, B. (1999) Modern Information Retrieval, New York, ACM Press.

Bakshi, K. and Karger, D. (2005) 'End-User Application Development for the Semantic Web' in Decker S. (ed) Proceedings of the ISWC 2005 Workshop on The Semantic Desktop - Next Generation Information Management & Collaboration Infrastructure, Galway, Ireland.

Bates, M.J. (1989) 'The Design of Browsing and Berrypicking Techniques', University of California at Los Angeles, Los Angeles

Belew, R. K. (2000) Finding Out About, Cambridge, Cambridge University Press.

Bunz, U. (2004) 'The Computer-Email-Web (CEW) Fluency Scale - Development and Validation', International Journal of Human-Computer Interaction, vol. 17, issue 4, pp. 479-506.

Chakrabarti, S. (2003) Mining the web, San Francisco, Morgan Kaufmann Publishers

Chandler. (n.d.) A next-generation Personal Information Manager [online], http://chandler.osafoundation.org/philosophy.php [Accessed 30 April, 2007].

Chaudhri, V., Cheyer, A., Guili, R., Jarrold, B., Myers K.L. and Niekrasz, J. (2006) 'A Case Study in Engineering a Knowledge Base for a Personal Assistant' in Semantic Desktop and Social Semantic Collaboration Workshop at the International Semantic Web Conference, Athens, USA.

Clarke, S. and Willett, P. (1997) 'Estimating the recall performance of search engines' *ASLIB Proceedings*, vol. 49 no. 7, pp. 184-189.

ClearForest. (n.d.). Gnosis [online], http://sws.clearforest.com/Blog/?page_id=32 [Accessed 23 March, 2007].

Cockburn, A., Karlson, A., and Bederson, B.B. (2006). A Review of Focus and Context Interfaces. College Park, USA, University of Maryland Human-Computer Interaction Lab.

Dervin, B. (2003) 'A Sense-Making Methodology primer: What is methodological about Sense-Making', Meeting of the International Communication Association, San Diego.

Dumais S.T. , Furnas, G.W. , Landauer, T.K., Deerwester, S. and Harshman, R. (1988) 'Using Latent Semantic Analysis To Improve Access To Textual Information' in proceedings of the Conference on Human Factors in Computing Systems, Washington DC, pp. 281--285.

Ellis, D. (1987) 'The derivation of a behavioural model for information retrieval system design', University of Sheffield.

Fernandez-Garcia, N., Sauermann, L., Sanchez L. and Bernardi, A. (2006) 'PIMO Population and Semantic Annotation for the Gnowsis Semantic Desktop' in Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop at the ISWC, Athens, USA.

Gifford, D. K., Jouvelot, P., Sheldon, M. A. and O'Toole, J. W. (1991) 'Semantic file systems' in Proceedings of the Thirteenth ACM Symposium on Operating Systems Principles, Pacific Grove, USA, pp. 16-25.

Gnowsis (n.d.). About [online], http://www.gnowsis.org/About/ [Accessed 23 August, 2007].

Johnson B. (2007) 'Google knows all with Web History', Guardian [online], http://www.guardian.co.uk/technology/2007/aug/02/googlewebhistory?gusrc=rss&feed=technology [Accessed 26 August, 2007].

Hackos, J. (1999) 'An Application of the Principles of Minimalism to the Design of Human-Computer Interfaces', Common Ground, vol. 9, pp. 17-22.

Horrocks I., Patel-Schneider P.F. and Van Harmelen.F. (2003). 'From SHIQ and RDF to OWL: The Making of a Web Ontology Language', Journal of Web Semantics, vol. 1, no. 1 (2003) pp. 7-26

Lawrence, S. and Giles, C.L. (1998) 'Searching the World Wide Web', Science, 280 (April), p. 98.

Marchionini, G. (2006) 'Exploratory Search: From Finding to Understanding', Communications of the ACM, vol. 49, no. 4., pp. 41-58.

Marczyk G., DeMatteo D. and Festinger D. (2005) Essentials of Research Design and Methodology, John Wiley & Sons, Inc.

Massachusett's Institute of Technology Haystack Project (n.d.). Haystack Client [online], http://haystack.csail.mit.edu [Accessed 17 March, 2007].

Meyers, M. (1997) 'Qualitative Research in Information Systems', MISQ Discovery, Management Information Systems Quarterly, June 1997.

Microsoft. (2004) 'Code Name WinFS: Revolutionary File Storage System Lets Users Search and Manage Files Based on Content', MSDN Magazine, (January).

Morita, T., Kura, T., Hidaka, T., Tanaka, A. and Kato, Y. (2006) 'Method of Retrieving a Web Browsing Experience using Semantic Periods' in Semantic Desktop and Social Semantic Collaboration Workshop at the International Semantic Web Conference, Athens, USA.

O'Day, V. and Jeffries R. (1993) 'Orienteering in an information landscape: how information seekers get from here to there' in Proceedings of the INTERCHI '93, Amsterdam, Netherlands.

Open University (2001) M873 User Interface Design and Evaluation: Evaluation, Milton Keynes, Open University.

Paulson, L. (2005) 'Building Rich Web Applications with AJAX', Computer, vol. 38, issue 10, pp. 14-17.

Pew. (2006) Pew Internet and American Life Project Surveys: Usage over time [online], http://www.pewinternet.org/trends/UsageOverTime.xls [Accessed 17 March 2007].

Quiroga, L.M. and Mostafa, J. (2002) 'An experiment in building profiles in information filtering: the role of context of user relevance feedback.' Information Processing & Management, vol. 38 no. 5 (September), pp. 671-694.

Salama, A., Samih, A., Ramadan, A. and Yousef K.M. (2006) GNU/Linux Semantic Storage System [online], http://www.glscube.org/downloads/glscube_design.pdf [Accessed 23 March, 2007].

Sharp J., Peters J. and Howard K. (2002) The Management of a Student Research Project, Third Edition, Open University.

Stone D., Jarrett C., Woodroffe M., and Minocha S. (2005) User Interface Design and Evaluation, Morgan Kaufmann.

Tang, C., Dwarkada S. and Xu Z. (2005) 'On scaling latent semantic indexing for large peer-to-peer systems' in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom pp. 112 - 121.

Voorhees E.M. (2005) 'The TREC robust retrieval track', ACM SIGIR Forum Trec Reports, vol. 39, no. 1, pp. 11-20.

Wilson, T.D. (1999) 'Models in Information Behaviour Research', Journal of Documentation, vol. 55, issue 3, pp. 249-70.

Wilson, T.D. (2000) 'Human Information behavior', Informing Science, vol. 3, issue 2, pp. 49-56.

Xiao H. and Cruz, I. (2005) 'A Multi-Ontology Approach for Personal Information Management' in 1st Workshop on The Semantic Desktop - Next Generation Personal Information Management and Collaboration Infrastructure, Galway, Ireland.

Xie, H. (2002) ‘Patterns between interactive intentions and information-seeking strategies’, Information Processing & Management, vol. 38, no. 1, (January), pp. 55-77.

Index

Appendix A – Empirical results summary

Group	Hypothesis	Criteria	Source	Technique	Approach	Result	Conclusion
Task context	1. The application is more useful for activities that are performed both locally and remotely.	Average number of relevant local files returned by task performed.	Observ. sessions	Machine relevance	Chi-test T-test	More relevant results returned for work/study tasks. Few relevant results returned for personal tasks.	Supported
		Average relevance rating for returned local files.	Observ. sessions	User-rated relevance	T-test	Higher relevance for work, study and search tasks. Low relevance for personal tasks.	Supported
	2. Fewer results equates with lower use	Trend between user reports of relevance and low hits/small indexes.	Observ. sessions	Index size to user reported relevance	Least-squares	Slight trend between low relevance and low hits/small indexes	Weakly supported
		Trend between low hits or small index size and application use.	Logging data	Average file or folder quantity vs. number of visits	Least-squares	No trend between file/fold quantity and number of visits.	Un-supported
				Index size versus page loads	least squares	Moderate trend between index size and number of visits.	Somewhat supported
	3. There is a correlation between often visited pages and greater relevance or quantity of local file results	Commonly accessed pages result in higher average amounts of local file hits and higher average relevance of those local file hits.	Logging data	Page loads vs. local hits.	Least-squares	No trend	Un-supported
				Page loads vs. Lucene relevance	Least-squares	No trend	Un-supported

Group	Hypothesis	Criteria	Source	Technique	Approach	Result	Conclusion
User type	4. Users who spend more time organising their data in many folders have more relevant results	Trend of greater use of the application by users with an average greater number of folders returned.	Logging data	Use analysis	Least-squares	No trend between greater use and greater folder quantity.	Un-supported
		Trend of relevance indications by users with a higher folder/file ratio.	Observ. sessions	User-rated relevance	Least-squares	No trend between relevance and high folder/file ratio.	Un-supported
	5. For some users, folders play an important role in interpreting results	Folder view selection frequency.	Logging data	Folder view selection vs. all page loads.	Simple frequency	Feature almost unused.	Un-supported
		Folder view toggle number of hits returned versus average number of hits returned		Folder view toggle hits returned vs. average hits returned	T-test	Difference in result set size, is larger or smaller.	Supported
		Users indicate they prefer folder view.	Observ. sessions	User-indicated preference	Least-squares	25% prefer folder view. Low correlation to small and large result sets.	Weakly supported
	6. Users recognise files by their filenames	Is metadata necessary in order to recognise results and contents?	Observ. Sessions	Recognise result set by filenames	Least-squares	Users recognize files by filename.	Supported

Group	Hypothesis	Criteria	Source	Technique	Approach	Result	Conclusion
Relevance factors	7. Only the most relevant results are at all relevant	Relative frequency of use of the "next 50" results button.	Logging data	Average number of hits when the next / previous feature used compared to average hits.	T-test	Feature almost unused.	Weakly supported
		Average relevance when the system was used in some way	Logging data	Lucene relevance versus use	Chi-test	Local file relevance more than double when system used.	Supported
		Higher machine-rated results actually more relevant.	Observ. Sessions	User-rated relevance	Chi-test	Users agreed with machine-relevance ratings.	Supported
	8. The more key phrases extracted, the lesser the relevance of the results	Trend between greater number of key phrases or page length and lower average relevance.	Logging data	Key phrase quantity vs. average relevance	Least-squares	Moderate trend between less key phrases and higher relevance.	Somewhat supported
				Page length vs. average relevance	Least-squares	No trend	Un-supported
				Page length vs. key phrase quantity	Least-squares	No trend	Un-supported
		User relevance rating compared with key phrase quantity and page length.	Observ. Sessions	Key phrase quantity vs. average relevance	Least-squares	No trend	Un-supported
				Page length vs. average relevance	Least-squares	No trend	Un-supported
	9. Frequently occurring key phrases are associated to more relevant results	Trend between common key phrases and actual system use.	Logging data	Common key phrases to average relevance.	Least-squares	No trend.	Un-supported

Group	Hypothesis	Criteria	Source	Technique	Approach	Result	Conclusion
Manual intervention	10. Filtering is a useful selection mechanism to increase relevance	Correlation between result set size and use of filtering functions.	Logging data	Use analysis	T-test	No correlation between result set size and use of filtering.	Un-supported
		Higher or lower average relevance ratings for filtered documents compared to relevance when viewing all documents.	Observ. sessions	User-rated relevance	Chi-test, T-test	Users found results more relevant when filtered.	Supported
	11. Key phrases extracted from title and metadata create more relevant results	How often users employed content, title or metadata to generate key phrases and the local file result relevance generated.	Logging data	Extraction method vs. average relevance.	Simple mean	Title generates high relevance Metadata generates similar relevance.	Partially supported
		User indications of relevance during observations sessions when asked to switch between content, title or metadata.	Observ. sessions	User indicated relevance comparing extraction techniques	Simple mean	Metadata slightly better. Title slightly worse.	Partially supported
	12. Query expansion using automatic relevance feedback is not very useful	How often users employ query expansion.	Logging data	Frequency of use	Frequency	Feature almost unused.	Weakly supported
		User indications of query expansion relevance.	Observ. sessions	Comparative relevance before and after query expansion using all key phrases.	Simple mean	Overwhelmingly makes results worse.	Supported
				Comparative relevance before and after query expansion using a single key phrase.	Simple mean	Makes results somewhat worse	Supported

Appendix B – Computer-Email-Web Fluency Survey

This questionnaire was designed by Ulla Bunz to assess computer literacy using a Computer-Email-Web (CEW) Fluency Scale (2004). There are about 70 multiple-choice questions. It should take about 5-10 minutes of your time. If a question is not applicable, leave it blank.

1. Please select your gender:

Male	64.3%	(9)
Female	35.7%	(5)
TOTAL	**100.0%**	**14**

2. What year were you born?

#	Response
1	1954
1	1964
2	1965
1	1966
1	1967
2	1969
1	1974
1	1975
2	1976
1	1977
1	1992

3. How long have you been using Internet (including using email, gopher, ftp, etc.)?

#	Response
2	10
2	11
2	12
2	14
1	15
1	16
1	17
1	20
1	5
1	9

4. How many computer classes, courses, or seminars have you attended throughout your lifetime?

#	Response
1	1
1	10

1	15
2	2
1	20
2	25
1	4
1	5
1	6
1	7

5. How frequently do you access the World Wide Web (WWW) from the home?

Daily	64.3%	(9)
Weekly	21.4%	(3)
Monthly		(0)
Once a month	7.1%	(1)
Never	7.1%	(1)
TOTAL	**100.0%**	**14**

6. How frequently do you access the World Wide Web (WWW) from work?

Daily	92.9%	(13)
Weekly		(0)
Monthly		(0)
Once a month		(0)
Never	7.1%	(1)
TOTAL	**100.0%**	**14**

7. How frequently do you access the World Wide Web (WWW) from school?

Daily	14.3%	(2)
Weekly	7.1%	(1)
Monthly		(0)
Once a month		(0)
Never	64.3%	(9)
TOTAL	**85.7%**	**14**

8. How frequently do you access the World Wide Web (WWW) from public terminals?

Daily		(0)
Weekly		(0)
Monthly	7.1%	(1)
Once a month	35.7%	(5)
Never	50.0%	(7)
TOTAL	**92.9%**	**14**

9. How frequently do you access the World Wide Web (WWW) from other locations?

Daily		(0)
Weekly	14.3%	(2)
Monthly	21.4%	(3)
Once a month	28.6%	(4)
Never	28.6%	(4)
TOTAL	**92.9%**	**14**

10. On average, how often do you use a WWW browser? By this, we mean using your browser for a specific set of tasks or activities. We do not mean how many times you launch

your browser per day.

More than 9 times/day	71.4%	(10)
5 to 8 times/day	7.1%	(1)
1 to 4 times/day	21.4%	(3)
A few times a week		(0)
Once a week		(0)
Once a month		(0)
TOTAL	**100.0%**	**14**

11. On average, how many hours a week do you use a WWW browser?

0 to 1 hours/week		(0)
2 to 4 hours/week		(0)
5 to 6 hours/week		(0)
7 to 9 hours/week	14.3%	(2)
10 to 20 hours/week	28.6%	(4)
21 to 40 hours/week	28.6%	(4)
Over 40 hours/week	28.6%	(4)
TOTAL	**100.0%**	**14**

12. What do you primarily use the Web for? (Please check all that apply.)

Education	85.7%	(12)
Shopping/gathering product information	64.3%	(9)
Entertainment	71.4%	(10)
Work/Business	85.7%	(12)
Communication with others (not including email)	85.7%	(12)
Gathering information for personal needs	92.9%	(13)
Wasting time	35.7%	(5)

13. Which of the following have you done? (Please check all that apply.)

Ordered a product/service by filling out an online form	92.9%	(13)
Made a purchase online for more than $100	78.6%	(11)
Created a web page	85.7%	(12)
Customized a web page for yourself (e.g. MyYahoo, CNN Custom News)	92.9%	(13)
Changed your browser's "startup" or "home" page	85.7%	(12)
Changed your "cookie" preferences	78.6%	(11)
Participated in an online chat or discussion (not including email)	85.7%	(12)
Listened to a radio broadcast online	100.0%	(14)
Made a telephone call online	78.6%	(11)
Used a nationwide online directory to find an address or telephone number	92.9%	(13)

Taken a seminar or class about the Web or Internet	35.7%	(5)
Bought a book to learn more about the Web or Internet	28.6%	(4)

14. How comfortable do you feel using computers, in general?

Very comfortable	71.4%	(10)
Somewhat comfortable	21.4%	(3)
Neither comfortable nor uncomfortable		(0)
Somewhat uncomfortable	7.1%	(1)
Very uncomfortable		(0)
TOTAL	**100.0%**	**14**

15. How comfortable to you feel using the Internet?

Very comfortable	100.0%	(14)
Somewhat comfortable		(0)
Neither comfortable nor uncomfortable		(0)
Somewhat uncomfortable		(0)
Very uncomfortable		(0)
TOTAL	**100.0%**	**14**

16. How satisfied are you with your current skills for using the Internet?

Very satisfied – I can do everything that I want to do	28.6%	(4)
Somewhat satisfied – I can do most things I want to do	71.4%	(10)
Neither satisfied nor unsatisfied		(0)
Somewhat unsatisfied – I can't do so many things I would like to do		(0)
Very unsatisfied – I can't do most things I would like to do		(0)
TOTAL	**100.0%**	**14**

---

17. I can print a document.

very well	71.4%	(10)
well	28.6%	(4)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

18. I can open a web address directly.

very well	85.7%	(12)
well	7.1%	(1)
okay	7.1%	(1)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

19. I can use search engines such as Yahoo or Alta Vista.

very well	71.4%	(10)

well	28.6%	(4)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

20. I can use "save as" when appropriate.

very well	85.7%	(12)
well	14.3%	(2)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

21. I can use the "reply" and "forward" features for email.

very well	92.9%	(13)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

22. I can save text contents off web pages to a disk

very well	85.7%	(12)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all	7.1%	(1)
TOTAL	**100.0%**	**14**

23. I can identify the host server from the web address.

very well	57.1%	(8)
well	21.4%	(3)
okay	7.1%	(1)
not so well	7.1%	(1)
not at all	7.1%	(1)
TOTAL	**100.0%**	**14**

24. I can read new mail messages.

very well	92.9%	(13)
well		(0)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**92.9%**	**14**

25. I can delete read email

very well	100.0%	(14)
well		(0)
okay		(0)

not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

26. I can send an email message.

very well	100.0%	(14)
well		(0)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

27. I can save images off web pages to a disk.

very well	78.6%	(11)
well	21.4%	(3)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

28. I can open an email program.

very well	100.0%	(14)
well		(0)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

29. I can edit bookmarks.

very well	92.9%	(13)
well		(0)
okay		(0)
not so well	7.1%	(1)
not at all		(0)
TOTAL	**100.0%**	**14**

30. I can open a previously saved file from any drive/directory.

very well	85.7%	(12)
well	14.3%	(2)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

31. I can open a file attached to an email.

very well	92.9%	(13)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all		(0)

TOTAL	**100.0%**	**14**

32. I can restart a computer.

very well	92.9%	(13)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

33. I can begin a new document.

very well	92.9%	(13)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

34. I can use a browser such as Netscape or Explorer to navigate the World Wide Web.

very well	92.9%	(13)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

35. I can create a website.

very well	35.7%	(5)
well	28.6%	(4)
okay	14.3%	(2)
not so well	14.3%	(2)
not at all	7.1%	(1)
TOTAL	**100.0%**	**14**

36. I can switch a computer on.

very well	92.9%	(13)
well	7.1%	(1)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

37. I can use "back" and "forward" to move between pages.

very well	100.0%	(14)
well		(0)
okay		(0)
not so well		(0)
not at all		(0)
TOTAL	**100.0%**	**14**

---

38. I use a computer to save time on work that would take me longer otherwise.

very frequently	50.0%	(7)
frequently	28.6%	(4)
sometimes	21.4%	(3)
rarely		(0)
never		(0)
TOTAL	**100.0%**	**14**

39. I use a computer to create professional-looking work.

very frequently	50.0%	(7)
frequently	28.6%	(4)
sometimes	14.3%	(2)
rarely	7.1%	(1)
never		(0)
TOTAL	**100.0%**	**14**

40. I play games on a computer.

very frequently	7.1%	(1)
frequently		(0)
sometimes	35.7%	(5)
rarely	28.6%	(4)
never	28.6%	(4)
TOTAL	**100.0%**	**14**

41. I do work by hand even though it would be faster on a computer.

very frequently		(0)
frequently	7.1%	(1)
sometimes	50.0%	(7)
rarely	28.6%	(4)
never	14.3%	(2)
TOTAL	**100.0%**	**14**

42. I use a computer to fill free time.

very frequently		(0)
frequently	14.3%	(2)
sometimes	50.0%	(7)
rarely	28.6%	(4)
never	7.1%	(1)
TOTAL	**100.0%**	**14**

43. I use a computer to procrastinate from doing work.

very frequently		(0)
frequently	14.3%	(2)
sometimes	50.0%	(7)
rarely	28.6%	(4)
never	7.1%	(1)
TOTAL	**100.0%**	**14**

44. I lose track of time while using a computer.

very frequently	7.1%	(1)
frequently	42.9%	(6)

sometimes	28.6%	(4)
rarely	21.4%	(3)
never		(0)
TOTAL	**100.0%**	**14**

45. I do work by hand because it is faster than doing it on a computer.

very frequently		(0)
frequently	7.1%	(1)
sometimes	42.9%	(6)
rarely	28.6%	(4)
never	21.4%	(3)
TOTAL	**100.0%**	**14**

46. I do work by hand that would look better if I did it on a computer.

very frequently		(0)
frequently		(0)
sometimes	35.7%	(5)
rarely	21.4%	(3)
never	42.9%	(6)
TOTAL	**100.0%**	**14**

47. I use a computer to do higher-quality work than I could do otherwise.

very frequently	28.6%	(4)
frequently	64.3%	(9)
sometimes	7.1%	(1)
rarely		(0)
never		(0)
TOTAL	**100.0%**	**14**

48. I use a Network/Internet to meet new people.

very frequently	7.1%	(1)
frequently	7.1%	(1)
sometimes	7.1%	(1)
rarely	28.6%	(4)
never	50.0%	(7)
TOTAL	**100.0%**	**14**

49. I use a Network/Internet to talk to people I see regularly in person.

very frequently	42.9%	(6)
frequently	21.4%	(3)
sometimes	14.3%	(2)
rarely	21.4%	(3)
never		(0)
TOTAL	**100.0%**	**14**

50. I use a Network/Internet to shop/look at products I would like to buy.

very frequently	21.4%	(3)
frequently	21.4%	(3)
sometimes	57.1%	(8)
rarely		(0)

never		(0)
TOTAL	**100.0%**	**14**

51. I spend time learning about the computer or Network/Internet itself.

very frequently	7.1%	(1)
frequently	21.4%	(3)
sometimes	57.1%	(8)
rarely	7.1%	(1)
never	7.1%	(1)
TOTAL	**100.0%**	**14**

52. I shop for computer hardware or software by going to stores or looking at catalogs.

very frequently		(0)
frequently	14.3%	(2)
sometimes	28.6%	(4)
rarely	42.9%	(6)
never	14.3%	(2)
TOTAL	**100.0%**	**14**

53. I spend time downloading and/or installing software.

very frequently	14.3%	(2)
frequently	50.0%	(7)
sometimes	14.3%	(2)
rarely	14.3%	(2)
never	7.1%	(1)
TOTAL	**100.0%**	**14**

54. I use a Network/Internet to keep in touch with friends and family who are far away.

very frequently	42.9%	(6)
frequently	42.9%	(6)
sometimes	14.3%	(2)
rarely		(0)
never		(0)
TOTAL	**100.0%**	**14**

55. I spend time configuring the computer to look and act as I want it to.

very frequently	35.7%	(5)
frequently	35.7%	(5)
sometimes	7.1%	(1)
rarely	21.4%	(3)
never		(0)
TOTAL	**100.0%**	**14**

56. How frequently do you access the Internet to access newsgroups?

Daily	35.7%	(5)
Weekly	28.6%	(4)
Monthly	7.1%	(1)
Once a month		(0)
Never	28.6%	(4)
TOTAL	**100.0%**	**14**

57. How frequently do you access the Internet to access online news?

Daily	85.7%	(12)
Weekly	7.1%	(1)
Monthly		(0)
Once a month	7.1%	(1)
Never		(0)
TOTAL	**100.0%**	**14**

58. How frequently do you access the Internet to access information about commercial products/services?

Daily	7.1%	(1)
Weekly	42.9%	(6)
Monthly	28.6%	(4)
Once a month	21.4%	(3)
Never		(0)
TOTAL	**100.0%**	**14**

59. How frequently do you access the Internet to purchase commercial products/services?

Daily		(0)
Weekly	14.3%	(2)
Monthly	35.7%	(5)
Once a month	42.9%	(6)
Never	7.1%	(1)
TOTAL	**100.0%**	**14**

60. How frequently do you access the Internet to access reference materials?

Daily	50.0%	(7)
Weekly	28.6%	(4)
Monthly	21.4%	(3)
Once a month		(0)
Never		(0)
TOTAL	**100.0%**	**14**

61. How frequently do you access the Internet to access research reports & projects?

Daily	28.6%	(4)
Weekly	42.9%	(6)
Monthly	14.3%	(2)
Once a month	7.1%	(1)
Never	7.1%	(1)
TOTAL	**100.0%**	**14**

62. How frequently do you access the Internet to access financial information?

Daily	7.1%	(1)
Weekly	28.6%	(4)
Monthly	14.3%	(2)
Once a month	21.4%	(3)
Never	28.6%	(4)
TOTAL	**100.0%**	**14**

63. How frequently do you access the Internet to access health/medical information?

Daily		(0)

Weekly	21.4%	(3)
Monthly	28.6%	(4)
Once a month	28.6%	(4)
Never	21.4%	(3)
TOTAL	**100.0%**	**14**

64. How frequently do you access the Internet to access online chat groups?

Daily	7.1%	(1)
Weekly	14.3%	(2)
Monthly	7.1%	(1)
Once a month	35.7%	(5)
Never	35.7%	(5)
TOTAL	**100.0%**	**14**

65. How frequently do you access the Internet to access online job listings?

Daily		(0)
Weekly	14.3%	(2)
Monthly	14.3%	(2)
Once a month	57.1%	(8)
Never	14.3%	(2)
TOTAL	**100.0%**	**14**

66. How frequently do you access the Internet to access online home/rental listings?

Daily		(0)
Weekly	21.4%	(3)
Monthly	7.1%	(1)
Once a month	35.7%	(5)
Never	35.7%	(5)
TOTAL	**100.0%**	**14**

67. How frequently do you access the Internet to access online telephone listings?

Daily		(0)
Weekly	14.3%	(2)
Monthly		(0)
Once a month	57.1%	(8)
Never	28.6%	(4)
TOTAL	**100.0%**	**14**

68. How frequently do you access the Internet to access online maps?

Daily	14.3%	(2)
Weekly	35.7%	(5)
Monthly	28.6%	(4)
Once a month	14.3%	(2)
Never		(0)
TOTAL	**92.9%**	**14**

69. How frequently to you use the Web instead of watching TV?

Daily	28.6%	(4)
Weekly	42.9%	(6)
Monthly		(0)

Once a month	7.1%	(1)
Never	14.3%	(2)
TOTAL	**92.9%**	**14**

70. How frequently to you use the Web instead of talking on the phone?

Daily	21.4%	(3)
Weekly	57.1%	(8)
Monthly	7.1%	(1)
Once a month	7.1%	(1)
Never	7.1%	(1)
TOTAL	**100.0%**	**14**

71. How frequently to you use the Web instead of sleeping?

Daily		(0)
Weekly	35.7%	(5)
Monthly		(0)
Once a month	21.4%	(3)
Never	42.9%	(6)
TOTAL	**100.0%**	**14**

72. How frequently to you use the Web instead of exercising?

Daily	28.6%	(4)
Weekly	14.3%	(2)
Monthly	7.1%	(1)
Once a month		(0)
Never	42.9%	(6)
TOTAL	**92.9%**	**14**

73. How frequently to you use the Web instead of reading books/magazines/newspapers?

Daily	28.6%	(4)
Weekly	28.6%	(4)
Monthly	21.4%	(3)
Once a month	14.3%	(2)
Never	7.1%	(1)
TOTAL	**100.0%**	**14**

74. How frequently to you use the Web instead of going to the movies?

Daily	7.1%	(1)
Weekly	21.4%	(3)
Monthly	7.1%	(1)
Once a month	21.4%	(3)
Never	42.9%	(6)
TOTAL	**100.0%**	**14**

75. How frequently to you use the Web instead of going out/socializing?

Daily	7.1%	(1)
Weekly	14.3%	(2)
Monthly	28.6%	(4)
Once a month	7.1%	(1)
Never	42.9%	(6)

TOTAL	**100.0%**	**14**

76. How frequently to you use the Web instead of doing household work?

Daily	21.4%	(3)
Weekly	14.3%	(2)
Monthly	14.3%	(2)
Once a month	7.1%	(1)
Never	42.9%	(6)
TOTAL	**100.0%**	**14**

Appendix C – Observation session script

Dear John Doe,

Thank you for agreeing to be part of the user testing of Context Explorer. I'll be reading from a script to insure that each user session is conducted in a similar way.

The purpose of this user session is to explore the ways in which people search and locate files on their computer and pages on the internet to try and better understand the connection between the two.

This user session is not meant to test or grade your skills as a computer user in any way so you should not feel embarrassed or under pressure to perform or answer any questions in a particular way.

This session will consist of the following activities:

- A survey on computer use
- A set of tasks to evaluate the utility of the Context Explorer application.

Let's go ahead and begin with the the survey. There was a link in the invitation letter for the first survey. Did you manage to complete it?

[If yes, continue to installation else access the Ulla Bunz computer survey:]

If a question doesn't apply, leave it blank.:
`http://www.organicdesktop.com/survey/public/survey.php?name=computer_literacy&username`

Now that we've completed the survey we need to make sure Context Explorer is installed on your computer and that the Regain indexer has indexed your documents.

[Verify installation]

<u>If not installed:</u>
[Install from web or data stick in the following order: Firefox, Context Explorer (Regain portion), configure file folders, install Sidebar. Note profile folder]

<u>If installed:</u>
[Verify version 1.3. Check file folders. Note profile folder]

Now that Context Explorer has been configured, have you used Context Explorer before and if so are you comfortable operating it?

[If no, give training session script and record any questions, hesitations, problems or misunderstandings.]

I will ask you to perform a series of tasks. During each task we will follow a set of steps and I'll record some information.

[Follow task script and complete logbook for each task as indicated]

That's it. Thank you so much for your time.

Appendix D – Task script

Remember, this user session is not testing your computing abilities. It is normal if you have difficulty using some of the features. Asking for help is part of the session.

For each task, wherever it says *evaluate* we will follow a series of defined steps.

1. **Task work remote**
    a. Open in your Firefox browser an unencrypted (not bank or pay pal) web page that is strongly associated to your everyday work or studies and *evaluate*.

2. **Task work local**
    a. Using the standard file explorer choose a document-like file on your computer strongly associated to your everyday work or studies.
    b. Open in your Firefox browser an unencrypted (not bank or pay pal) web page that you believe is related to the local file you opened and *evaluate*.

3. **Task personal**
    a. Open in your Firefox browser an unencrypted (not bank or pay pal) web page that you go to commonly for personal reasons (not for work or studies) and *evaluate*.

4. **Task search**
    a. Come up with a key phrase or key phrase strongly associated to your everyday work or studies.
    b. Using a web search engine of your choice, query using the key phrase or phrase and using the result page, evaluate.

Appendix E – Training script

Context Explorer is a fairly simple application. In the background while you browse the web it extracts the key phrases from your web pages and uses them to query your local computer for matching resources. It dynamically displays those matches as you browse, giving you a constantly updated view of the files on your computer that match the pages you browse.

Let's have a quick look at its functions. ***[Open CE if it is not already open]***

Web page extraction type

This section controls what part of the web page is used for extracting the key phrases. All is everything, Title is just the title you see in the top bar of the browser, metadata is the bibliographic subject data that sometimes is attached to a page, and content is the part of the page you see without the title or metadata.

The "irrelevant keys" button on the right shows you any key phrases found on a page that were not associated to anything on your computer.

The "expand query" button on the right allows you to research your computer using key phrases found in the current set of most relevant documents on your computer.

Relevant extracted key phrases

This section shows the key phrases that have been extracted from the current web page together with the number of files on your computer that were linked to that key phrase. Clicking on any of them restricts the results to just that key phrase.
Found local document types
This section shows you the kind of documents that were found. You can filter the results by clicking on them.

Local results

This section shows you the files that were found on your computer that are related to the current web page. The file and folder icons allow you to switch between viewing the related files, or the folder paths in which the files were found.
The black results are more relevant, grey results less relevant. Hovering over them will give you a summary of the file contents. Clicking on a file allows you to open it, or open the folder it was found in.

If there are more than 50 results you can jump forward or backward to view the rest of the results.

Appendix F – Logbook

User			
Place			
Role of computer			
Date		**Time**	
Firefox profile folder		**Operating system**	
Is only user?		**Single user index?**	

1. User require training?

True	False

2. Note any confusion, hesitation, difficulty using application:

3. Time needed to master the application: __________

4. Index size ___________

Task: ________________
URL:

Count the number of folders (relevant)		
Write down the number of hits (all, relevant)		
Write down the number of relevant key phrases		
Write down the number of document types		
Write down number of words on page		

1. How relevant is the overall contextualization (web page, key phrases & results) at first glance?

Not at all	A little	Somewhat	Very	Excellent

2. Look at just the key phrases.
   a. Rate their relevance to the web page:

Not at all	A little	Somewhat	Very	Excellent	Percent:	

   b. Rate their relevance to the black file results:

Not at all	A little	Somewhat	Very	Excellent	Percent:	

   c. Rate their relevance to the activities you normally associate to this page:

N/A	Not at all	A little	Somewhat	Very	Excellent

3. Look at just the file results.
   a. Rate the relevance of the black file results to the current web page:

Not at all	A little	Somewhat	Very	Excellent	Percent:	

   b. Rate the relevance of the black file results to the activities you normally associate to this page:

Not at all	A little	Somewhat	Very	Excellent	Percent:	

   c. Compared to the web page, are the black file results more relevant than the grey file results?

Much worse	Worse	Same	Better	Much better

4. Do you prefer the file or folder view for this result set?

No preference	File view	Folder view

5. For the black file results, do you recognize them by their filenames without hovering over them and looking at their metadata?

None	A few	Some	Many	All

6. Look at the metadata for the first three file results, how helpful is it in recognising or interpreting the file contents?

Not at all	A little	Somewhat	Very	Excellent

7. Switch between the *Content, Title* and *Metadata* content extraction methods. For each, rate the change in the relevance of the black file results to the web page:
   a. Content

Error	N/A	Much worse	Worse	Same	Better	Much better

b. Title

Error	N/A	Much worse	Worse	Same	Better	Much better

c. Metadata

Error	N/A	Much worse	Worse	Same	Better	Much better

8. Return to the *All* view and switch between as many as five document types. Rate the relevance of the black file results (if any) to the web page:

   a. Document type:

N/A	Not at all	A little	Somewhat	Very	Excellent

   b. Document type:

N/A	Not at all	A little	Somewhat	Very	Excellent

   c. Document type:

N/A	Not at all	A little	Somewhat	Very	Excellent

   d. Document type:

N/A	Not at all	A little	Somewhat	Very	Excellent

   e. Document type:

N/A	Not at all	A little	Somewhat	Very	Excellent

9. Switch back to the *All* document types view, click the "expand query" button and rate the change in the relevance of the black file results to the web page:

Much worse	Worse	Same	Better	Much better

10. Refresh by clicking the *Content* view and then the *All* view again. Click the key phrase that is most relevant to the activities you normally associate to this page and rate the change in the relevance of the black file results to the web page:

Much worse	Worse	Same	Better	Much better

11. Click "expand query" again and rate the change *from the previous step* in the relevance of the black file results to the web page:

Much worse	Worse	Same	Better	Much better

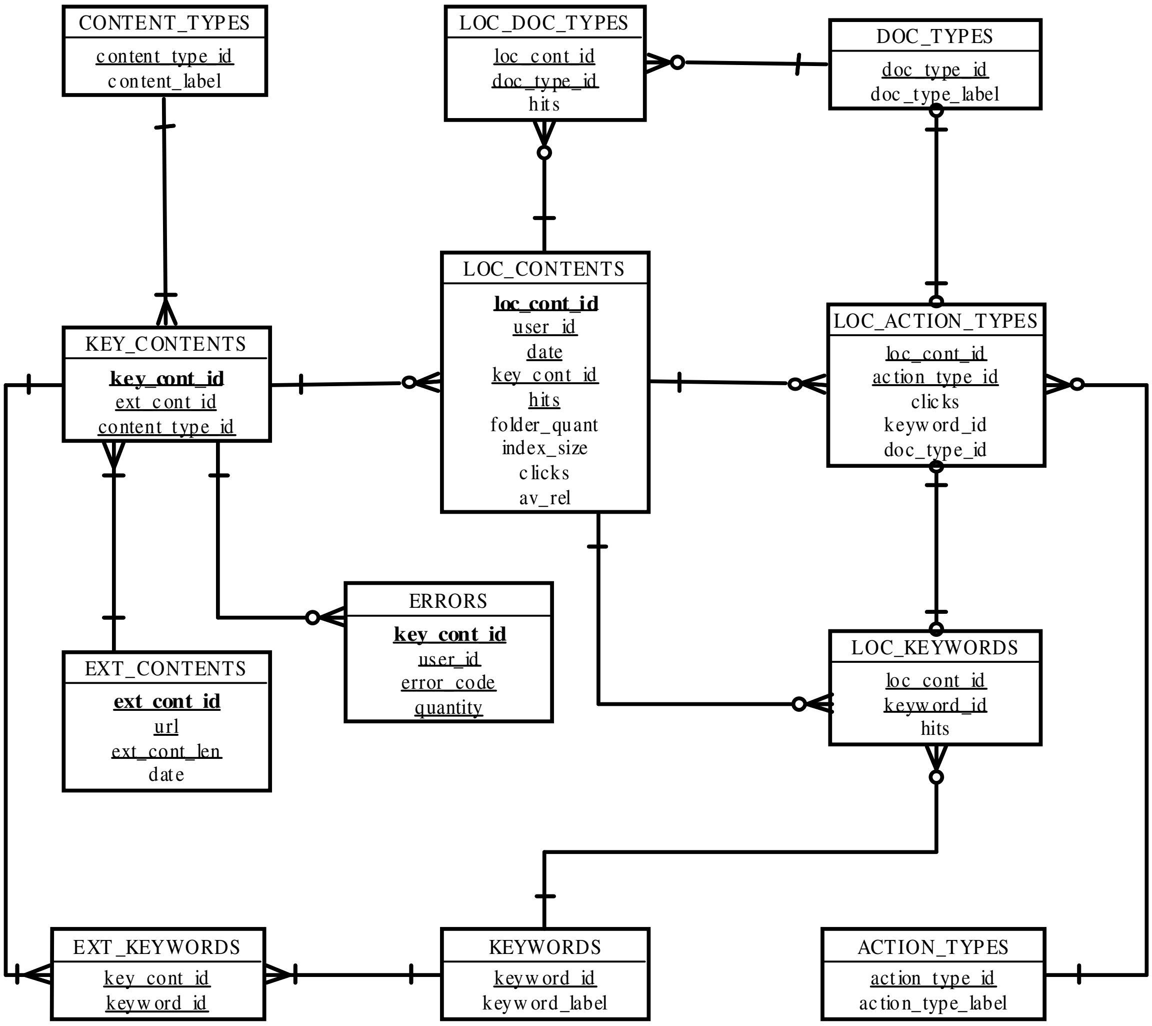
12. General comments on this activity:

Appendix G – Application screenshot

Context Explorer is designed to bridge the gap between what you're surfing on the web, and what you have on your computer. It's pretty simple: as you browse the web, it automatically scrapes the content from your pages and extracts the keywords they contain. It then uses those keywords to query your local system, using the index created by the Regain indexer installed as part of the setup.

If it's not working, the problem is almost always that the Regain portion of Context Explorer is not running, not configured, or not installed. You should see a blue R in your taskbar. If you don't see it, check to see if Context Explorer is in your startup menu and put it there if it's not. If you see the blue R but it still doesn't work, right-click the R and select "preferences". Make sure you have at least one directory listed. If you don't have a blue R and you don't have any Context Explorer menu items, you need to re-install!

Appendix H – Logging database design



ERRORS
1 Regain timed out
2 Regain error
3 Not enough page content
4 Page encrypted
5 Yahoo timed out
6 Yahoo error
7 no keywords extracted
8 no related resources
9 Logging timed out ext_cont
10 Logging error ext_cont
11 Logging timed out loc_cont
12 Logging error loc_cont
13 Logging timed out action
14 Logging error action

CONTENT_TYPES	
0 = all	2 = title
1 = metadata	3 = content

ACTION_TYPES	
0 select_keyword	11 view_folders
1 select_doctype	12 view_files
2 view_irrelevant_keywords	13 minimize
3 select_expand_query	14 maximize
4 view_next	15 select metadata
5 view_previous	16 select title
6 view_result_metadata	17 select content
7 open_file	18 select all
8 open_folder	19 open app
9 select_all_keywords	20 close app
10 select_all_doctypes	

Figure 19 - Context Explorer logging database

CONTENT_TYPES

Contains a list of the four content choices for keyword selection: all, title, content, metadata
content_type_id
The auto-generated primary key.
content_label
The name

KEY_CONTENTS

This table is the join of a web page and an extraction type, thus it represents a certain set of keywords consumed by the local indexer.
key_cont_id
The auto-generated primary key.
ext_cont_id
Foreign key, the web page.
content_type_id
Foreign key, The type of page extraction.

LOC_ACTION_TYPES

A table that associates user actions to current resources.
loc_cont_id
The set of current results.
action_type_id
The id of the user action
clicks
The number of times this user action has been recorded for these current results
keyword_id
If a keyword was clicked to refine results, its ID is stored.
doc_type_id
If a doc_type was clicked to refine results its ID is stored.

EXT_CONTENTS

The representation of a web page. Held in a separate table since multiple users might access the same URL. Its keys consist of the URL and the page length
ext_cont_id
The auto-generated primary key.
url
The page URL.
ext_cont_len
The page length
date
The date retrieved

LOC_CONTENTS

Represents a set of results on a user's local system linked to the keywords generated by a particular web page/content type.
loc_cont_id
The auto-generated primary key.
user_id
Combo key, the profile folder name from the user's Firefox installation.
key_cont_id
Foreign key, the page/content type combo to which this set of results is linked.
hits
Combo key, the number of hits generated by the association to the web page. If it changes, we know resources have been added/removed so we make a new entry.
date
The date.
folder_quant
How many relevant folders were returned.
index_size
The size of the index on the user's computer.
clicks
the number of times this content set has been accessed
av_rel
the average relevance for this result set

LOC_KEYWORDS

A join table that matches the keywords from the current content to the current set of results.
loc_cont_id
The set of current results.
keyword_id
A keyword.
hits
The number of hits associated to that keyword.

ERRORS

key_cont_id
The page/content combo that generated the error
user_id
The user id.
error_code
The error type
quantity
The number of times it has happened for this page/content/user combo

LOC_DOC_TYPES

loc_cont_id
The set of current results.
doc_type_id
A filetype extension found as a result
hits
The number of that doctype found for the current results.

DOC_TYPES

A list of mime-types and labels for the resource types found on a user's system.
doc_type_id
The filename extension.
doc_type_label
The mime-type

ACTION_TYPES

action_type_id
The auto-generated primary key.
action_type_label
The name of the action.

EXT_KEYWORDS

A join table that identifies the set of keywords produced by a piece of content.
key_cont_id
The content
keyword_id
The keyword id

KEYWORDS

A set of keyphrases that have been returned by various pages
keyword_id
The auto-generated primary key.
keyword_label
The keyword or phrase.

Figure 20 - Context Explorer logging database description

Appendix I – Source code and collected data

The source code and collected data for the project is attached to this thesis on a compact disc. It provides all the code necessary to repeat the thesis research. It includes the following:

- Context Explorer source code including:
    - Javascript
    - XUL
    - Cascading Style Sheets
    - Images
    - Chrome manifest
    - Install.rdf
    - Contextexplorer.dtd
- Mochikit Ajax library
- Regain desktop indexer including:
    - Standard Regain installer
    - Modified JSP for Regain used by Context explorer
    - Modified Regain XML configuration containing English stop words
- MySql databases
    - Logging database
    - Survey database
    - User observation database
- PHP logging mechanism
    - PHP files handling the connection between instances of Context explorer and the logging database