Refinery of an internet-based search tool: exploring perceptions from Information Systems practitioners

Michael Chau*

Faculty of Business and Economics, School of Business,
The University of Hong Kong, Pokfulam, Hong Kong
Fax: +852 2858 5614 E-mail: mchau@business.hku.hk
Website: http://www.business.hku.hk/~mchau/
*Corresponding author

Ivy Chan

Department of Decision Sciences and Managerial Economics,
The Chinese University of Hong Kong,
Shatin, N.T., Hong Kong
Fax: +852 2603 5104 E-mail: ivychan@baf.msmail.cuhk.edu.hk

Abstract: In today's dynamic business environment, the capability to understand the needs and responses of stakeholders is critical, as such management can devise effective plans and course of actions for long-term strategies. The advancement of the internet and its related search tools, such as Google, has helped management to collect various information in order to keep abreast of time of business environment and their business communities. Yet, the management has encountered problem and uncertainty on the information quality such as content, currency, accuracy and presentation for decision making. Therefore, a new backlink search tool, namely, Redips was developed to serve such purpose. This tool has been tested by student subjects and received satisfactory feedback with its usefulness and functionality. However, in order to gain acceptance from the management (i.e., the end-users), the perceptions and expectations on the tool should not be overlooked. This study presents a wide range of perceptions from a group of Information Systems (IS) practitioners on the potential use of Redips.

Author: Please reduce abstract to not more than 100 words.

Keywords: web search agents; business intelligence; backlink search; web analysis; focus group study.

Reference to this paper should be made as follows: Chau, M. and Chan, I. (xxxx) 'Refinery of an internet-based search tool: exploring perceptions from Information Systems practitioners', *Int. J. Electronic Business*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Michael Chau is an Assistant Professor in the School of Business at the University of Hong Kong. He received a PhD Degree in Management Information Systems from the University of Arizona and a Bachelor Degree in Computer Science and Information Systems from the University of Hong Kong. His current research interests include information retrieval, web mining, data mining, knowledge management, electronic commerce, security informatics, and intelligence agents.

Ivy Chan is an Instructor at the Chinese University of Hong Kong. She received her PhD in Business Administration from the School of Business at the University of Hong Kong. Her research interests include knowledge management, information systems planning and organisational learning.

1 Introduction

The World Wide Web presents significant opportunities for business intelligence analysis as it can provide information about a company's external environment and its stakeholders. Traditional business intelligence analysis on the web has focused on the simple keyword searching. Recently, it has also been suggested that the incoming links, or backlinks, of a company's website (i.e., other web pages which have a hyperlink pointing to the company of interest) can provide important insights about the company's 'online community'. Analysis of these communities can provide useful signals for a company or information about its stakeholder groups, but the manual analysis process can be very time-consuming for business analysts and consultants. In this paper, we present a tool called Redips that integrates automatically backlink meta-searching and text mining techniques to facilitate users in performing such business intelligence on the web. We also report a focus group study involving IS practitioners to study the potential uses and user perception of the proposed tool. The rest of the paper is structured as follows. Section 2 reviews the importance of web communities in business intelligence analysis. Section 3 presents the proposed tool and its architecture. In Section 4, we discuss our methodology and the focus group study. The preliminary results are presented in Section 5. Finally, we conclude our study and suggest future research directions in Section 6.

2 Literature review

2.1 Web communities

The internet has many well-known explicitly defined communities – groups of individuals who share a common interest, together with the web pages most popular amongst them (Reid, 2003). The web communities consist of the following stakeholders of the firm: customers, suppliers, competitors, regulators, employees, educational institutions, court and legal institutions, financial institutions, stockholders, public-interest groups, labour unions, political parties, federal, state, local governments, etc. (Schermerhorn, 2001). The stakeholders listed here can be classified into two categories: explicit and implicit web communities.

Explicit communities are the groups that can be easily identified on the internet. Kumar et al. (1998) discussed the Porsche newsgroup as an example of explicit community of web users interested in Porsche Boxster cars. Such communities are often found in resource collections in web directories such as the Yahoo directory. Explicit communities are easy to be identified and analysts can simply use manual method to find a firm's explicit communities by browsing the firm's newsgroup or the category in which the firm belongs to in directories like Yahoo on the internet.

Implicit communities are relatively more difficult to be found using a manual browsing method. According to Kumar et al. (1998), implicit communities refer to the distributed, ad-hoc and random content-creation related to some common interests on the internet. These pages often have links to each other, but the common interests of implicit communities are sometimes too narrow and detailed for the resource pages or the directories to develop explicit listings for them. As a result, it is more difficult to find the implicit communities of a firm. In identifying the explicit and implicit communities of a firm, it is reasonable to assume that the content pages created by these communities would provide hypertext links back to the firm's homepage for reference (Reid, 2003). Therefore, in order to find a firm's online communities, it is necessary to find the web pages that have hyperlinks pointing to the firm's URL, i.e., the inbound links of the firm's website.

The identification of web communities, irrespective of explicit communities or implicit communities, is important to the strategic planning process. The strategic planning process consists of five steps, namely, mission and objectives, environment scanning, strategy formulation, strategy implementation and evaluation and control (Bradford et al., 1999). The extraction of web communities is classified to the environmental scanning step in the strategic planning process. The information would be used for the analysis of the firm's industry for evaluating entry barriers, suppliers, customers, substitute products and industry rivalry.

2.2 Web analysis tools

Many tools have been used to assist web-based business intelligence analysis. The simplest tool may be just a web browser like Internet Explorer. A browser is a client software program used for searching and viewing various kinds of information on the web. Using a manual browsing method, an analyst only needs to enter the URL of a firm or its stakeholders in the browser and then manually browse the information for further analysis.

This manual browsing method is common to analysts. It is simple as many people nowadays are experienced in internet surfing. Manual browsing also ensures the quality of the information collected. However, the process of manual browsing is very time-consuming and mentally exhausting. Data collection is the most time-consuming task in typical analysis projects, accounting for more than 30% of the total time spent (Prescott and Smith, 1991). It is not practical for analysts to go through the websites of all stakeholders of a company in detail. To make the problem worse, many web pages are updated weekly, daily or even hourly. It is almost impossible for analysts to manually collect the most updated versions of every web page for analysis.

To address these problems, web-based business intelligence tools have been developed to do more than simple browsing. In the following, we will review literature and existing tools in web-based business intelligence analysis. Based on their functionalities, the tools can be classified into three categories, namely, web searching, content analysis and visualisation. The discussion in the following will be based on this taxonomy.

2.2.1 Web search engines

Web search engines are the most popular way that people use to search for information on the web. Each engine has its own characteristics and employs its preferred algorithm in indexing and ranking web documents. For example, Google (www.google.com) and AltaVista (www.altavista.com) allow users to submit queries and present the search results in a ranked order, while Yahoo (www.yahoo.com) groups websites into categories, creating a hierarchical directory of a subset of the web. A web search engine usually consists of four main components: spiders, indexer, retrieval and ranking and user interface. Readers can refer to Brin and Page (1998) and Chau and Chen (2003) for a detailed technical description.

Another type of search engine is meta-search engine, such as MetaCrawler (www.metacrawler.com) and Dogpile (www.dogpile.com). These search engines do not keep their own indexes. When a search request is received, a meta-search engine connects to multiple popular search engines and integrates the results returned by these search engines. As each search engine covers different portion of the internet, meta-search engines are useful when the user needs to get as much of the internet as possible (Selberg and Etzioni, 1997; Chen et al., 2001).

In addition to general searching, analysts can also use *backlink searching* to research a firm's web communities that consist of the important stakeholders of the firm. Backlink searching can identify these communities as the stakeholders generally have on their web pages a hyperlink that point to the URL of the firm. Some general search engines also provide the feature of backlink searching. In these search engines, the indexer will, in addition to performing regular indexing, also index the links of each web page collected. The information on these links is stored into the search engine's database, so it is possible for users to search for all links that point to a given web page. One example is the Google search engine (www.google.com). Google allows users to use the reserved word 'link' as an operator in the query. The query 'link:siteURL' shows the users pages that point to a given URL (Google, 2005). For example, the query 'link:www.google.com' will return pages that contain a hyperlink to Google's home page. A special query capability using the query prefix 'link:', lists web pages that have links to the specified web page. Altavista (www.altavista.com) and MSN Search (search.msn.com) also have a similar feature and a similar 'link:' operator that finds pages with a link to a page with the specified URL.

Unlike general-purpose searching, no meta-search engines are available for searching backlinks in the current search engine market. A meta-backlink search engine may be able to improve retrieval performance just like general meta-search engines, but this has not been studied in previous research.

2.2.2 Web content analysis

After documents are retrieved from the web, indexing and text mining techniques are often applied to perform further analysis on the documents. Text mining, also known as text data mining (Hearst, 1997) or knowledge discovery from textual databases (Feldman and Dagan, 1995), refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents (Tan, 1999). Text mining serves as an extension of data mining or knowledge discovery from structured databases (Fayyad et al., 1996), which combines the knowledge from multiple

fields including information retrieval, textual information analysis, information extraction and information clustering.

Text mining tools help analysts to better understand the retrieved web document set from the internet, identify interesting web documents more effectively and gain a quick overview of the web documents' contents. This saves the manual browsing time of reading the entire set of web pages. Analysts only have to examine the categories, which are of the firm's interest.

As the documents on the web mainly contain textual contents, e.g., HTML documents or PDF documents, text mining and textual information analysis are often studied in internet-based analysis tools. Textual information analysis relies on the indexing of the source web documents. Many techniques of indexing the source documents and extracting key concepts from text have been proposed in the recent few years. One of the proven techniques is automatic indexing algorithm, which is regarded as effective as human indexing (Salton, 1986).

Automatic indexing algorithms can be based on either single words or phrases. Single word indexing allows users to search for documents that contain the search keywords and has widely adopted in information retrieval systems. The output is a vector of extracted words representing the documents of interest, based on each term's frequencies. Different from single word indexing, phrase indexing outputs a vector of extracted phrases to represent the documents of interest. An analysis tool Arizona Noun Phraser (AZNP) has been developed based on this concept (Tolle and Chen, 2000). The tool extracts all the noun phrases from each web document based on part-of-speech tagging and a set of linguistic rules.

The output of automatic indexing algorithms can often be used in further text mining analysis. Document classification or document clustering can be applied to the noun phrases in order to deduce patterns and relationship across documents and to derive firm-related knowledge in the analysis project of the firm. Document classification is one form of data analysis that can be built to categorise the documents into a predetermined set of document classes or concepts (Han and Kamber, 2001). Web documents are categorised into predefined classes in this approach. Since the classes or concepts are provided, the classification step is also known as supervised learning. This contrasts with unsupervised learning (or clustering), in which the classes are not known, and the number or set of classes to be learned also may not be known in advance. Clustering is the process of grouping objects into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. The classes or clusters would have a category label defined based on the keywords or phrases that appear in the web documents in that category. One of the popular clustering approaches for client-side internet analysis tool is Kohonen's Self-Organising Map (SOM). The SOM algorithm, which is a type of artificial neural network, classifies documents into categories that are determined during the clustering process (Kohonen, 1995). The algorithm clusters the retrieved documents into different regions and displays the results as a two-dimensional map. More details on its visualisation capabilities are discussed in the next subsection.

2.2.3 Web content visualisation

Visualisation tools are often used to display the document classification or document clustering results to users in an organised and meaningful way using certain graphical

representation. A graphical representation of the document clusters helps analysts and managers to better comprehend the returned documents, identify interesting documents more quickly, gain a quick overview of the documents' contents and acquire knowledge more effectively (Johnson, 1994). Furthermore, a graphical representation summarises the key results and shortens the time for users to digest the data, information, knowledge and intelligence in the documents.

Depending on the respective visualisation targets, visualisation tools can be classified into two categories. The first one is a category of tools that visualise the document attributes, e.g., document type, location, created date, title, document size, source, topic and author. The objective is to provide the users with additional information about the retrieved documents. Another category includes tools that utilise inter-document similarities to reduce the multidimensional document space to a two-dimensional or three-dimensional space (clusters) by aggregating similar documents under the same topic. The objective is to provide the users with a quick overview of the whole document collection. Cluster labels are decided based on the words or phrases written in the document collection.

Kohonen's self-organising map, as discussed earlier, is not only a clustering algorithm but also an example of visualisation tools based on inter-document similarities. This approach clusters documents into various topics that are automatically generated in the real time using neural network algorithms (Kohonen, 1995; Chen et al., 1996). Every document is assigned to its corresponding regions in a two-dimensional graphical map displayed to the user. Every region contains similar documents under the same topic while those regions with conceptually similar topics are located close to each other on the self-organising map.

Visualisation tools display graphical representation to the users and help them better understand the set of retrieved documents in a short amount of time, which is especially important in the today's fast-changing business world. Clearly, the visualisation tools improve the user experience and contribute as an important component in internet-based analysis tools. In spite of the benefits of visualisation tools, many tools discussed in this section are used only in the prototype research systems, and few is put on the commercial market for real business applications. Training and system assistance are needed to improve the effectiveness of clustering approaches and visualisation tools for internet-based analysis tools (Sutcliffe et al., 2000).

3 Redips

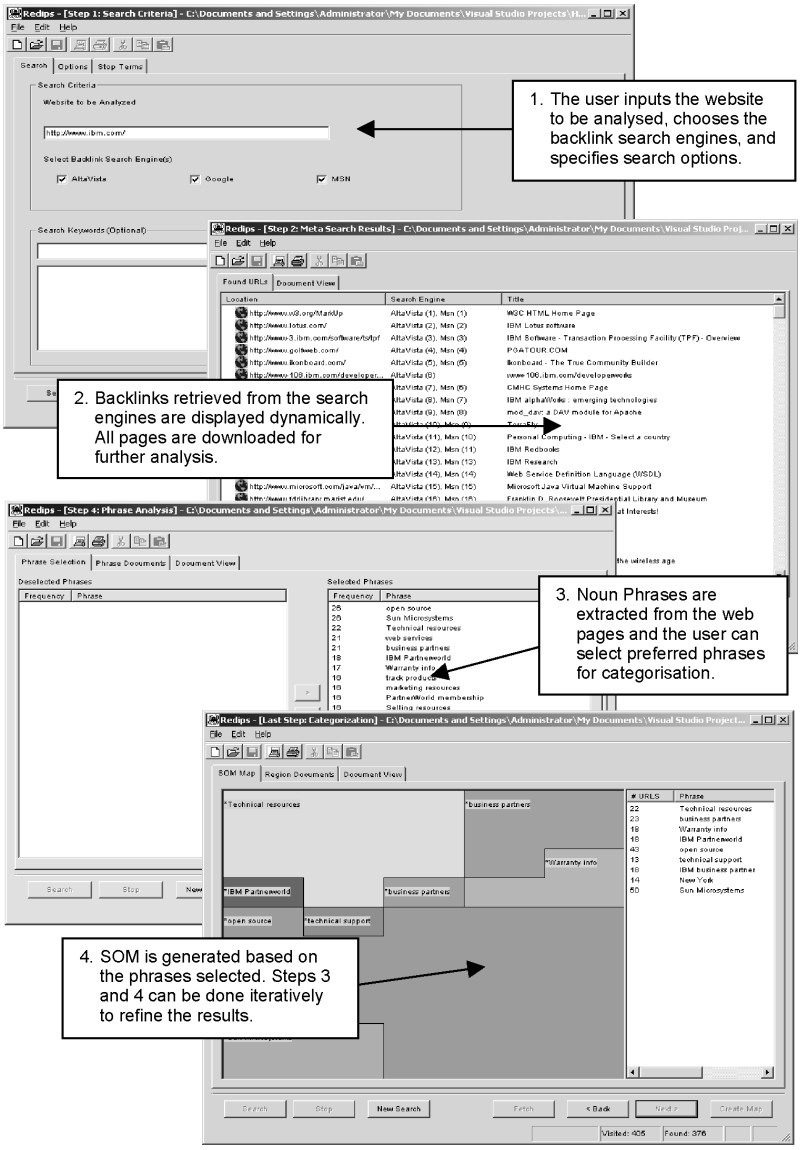
Our proposed system is called Redips. The term 'Redips' is the reverse spelling of the term 'Spider'. Traditional web spiders (crawlers) search the web by following outgoing links. However, in our proposed tool, we search based on the 'incoming links' of a page. These incoming links can help identify the web communities of the firm of interest.

Redips has been implemented based on the MetaSpider system developed in our previous research (Chen et al., 2001). The main modules include the user interface, spider, Arizona noun phraser, and self-organisation maps. User interface is the first point of contact between the user and the system. Spider fetched the URLs returned from those search engines. Arizona noun phraser is a natural language processing tool to do the key phrase extraction from internet text. Self-organisation maps visualise the concepts in a two-dimensional map. The technical details of the modules are discussed in another paper

(Chau et al., 2005). Here, we will present a search session with the tool from the user's perspective.

When using Redips, a user should first enter the website to be analysed and the backlink search engines to be included. A sample user session with Redips is shown in Figure 1. In this scenario, a business analyst wants to analyse the online community of the company IBM, so the analyst entered http://www.ibm.com/, the homepage of the IBM website, to the search query box. Optionally, the user can enter the keyword(s) to be included in the returned web pages. The user may also specify some other search options. In this step, the user can define the intelligent analysis objectives, e.g., the firm, information source, topic, in the analysis process.

Figure 1 Example of a user session with Redips



After submitting the search requests, the system collects from the different search engines the web pages that point to www.ibm.com. The results are displayed to the analyst as a list. In order to avoid causing confusion to the user, our system only groups URLs that are returned from more than one search engine, but does not perform any reranking of the search results in this step. The user can click on any URL and view the actual web page. Exploratory, preliminary research can be carried out to identify the firm's web communities in this step.

After browsing through the list of search results, the analyst wants to further analyse the relevant web pages in detail, so the analyst clicks on the *Fetch* button to command the system to download from the web the complete content of the web pages listed in the search results. The pages are displayed dynamically during the fetching process. The analyst can explore the results and read the content of any of the web pages collected. If the analyst wants to focus only on the valid pages, the display can be switched to the *Good URL List* to browse the filtered result. In this list, web pages that no longer exist (e.g., the page has been removed from the internet; the web server hosting the page reports a 404 'Page Not Found' error; or the web server does not respond with a specific timeframe) and those that do not contain the search keyword(s) will be excluded. Overall, this step automates the process of information collection from the web communities on the internet and allows the analyst to browse through the actual content of the web communities in detail.

After the complete content of all the web pages have been downloaded to the system, the analyst instructs the system to send the results to the Arizona Noun Phraser for further analysis. Noun phrases are extracted from the web pages and analysed. The frequency of appearance of each noun phrase is displayed and the analyst can browse the pages that contain any particular noun phrase by clicking on the phrase. In this step, the analyst can get a quick overview of the most frequent topics that appear in the web communities of IBM, e.g., 'open source', 'Sun Microsystems', 'technical resources' and 'web services'.

The analyst can then select what noun phrases are to be included in creating a categorisation map, known as the Self-Organising Map (SOM). In the map display, the web pages are categorised into different regions on the map based on their topic. The SOM help the analyst summarise and visualise the web communities in a graphical representation, which is useful in the business intelligence analysis process. In this example, a few web communities of IBM are identified on the map, e.g., IBM Partnerworld, Sun Microsystems, the open source community, and so on. More important communities occupy larger regions (e.g., open source), and similar communities are grouped close to each other on the map. Such information can help the analyst identify the web communities more easily. The analyst can also click on any category on the map and read the web pages in more detail for further analysis.

Through this simple example, we have shown how the system can help business analysts and management to gain a quick overview of the business environment faced by the organisation through web community analysis. The tool provides some automated web searching, web community analysis, text analysis and visualisation technologies that are useful for such purposes.

4 Research methodology

As discussed above, there are a number of new attributes embedded in Redips, which facilitate information search and add values to the management in strategic planning. Yet, the attributes of Redips perceived and developed by the system developers may be different from those of the business managers. In order to provide an effective and practical tool, the views and comments from IS practitioners who are rich in IS knowledge and currently engage in business planning, leadership and strategic analysis are examined. The current study adopts focus group study (as an effective research method) to dig in and explore various perceptions from the IS practitioners, as such to enrich our understanding of the practicability and usefulness of Redips.

Focus group discussion is a qualitative method to trigger and solicit extensive pool of opinions that capture the respective dimensions of the domains or topics to be addressed or investigated, particularly newly emerged and illuminative concepts (Krueger, 1998; Morgan, 1988). Generally speaking, focus group is a method conducive to communication through a two-hour back-and-forth discussion and interaction among a small group of individuals (usually 6–10 persons). A moderator is enlisted to facilitate the discussion (e.g., to ensure topics and issues are conversed, commented upon and incorporated). It is evident that focus group is commensurate with exploratory study to generate broad-based ideas that are relatively inaccessible using other research methods (Caldeira and Ward, 2002). In our current study, one focus group discussion was conducted. The group was composed of six practitioners (in their mid-1930s) with an average of 12 years of industry experience. They discussed on the search approaches, pros and cons of commonly used internet-based search tools. Thereafter, they engaged in a 30-minute hands-on of the Redips and explicated their opinions from various views. Large amount of broad-based ideas in participants' own words, while relatively inaccessible using other research methodologies can be obtained (Caldeira and Ward, 2002). The coding and analysis of the data, thereby can be used to develop thematic scheme to reveal the scope, details and multiple facets of the constructs or ideas in proposed work (Blackburn and Stokes, 2000; Stewart and Shamdasani, 2000).

The focus group study serves as the beginning stage of research. We intend to use a mix of techniques to make sense of the qualitative data collected from the practitioners. A research plan with three stages is detailed as follows:

- *Stage 1*: In order to understand how users perceive and evaluate new systems, a reference to existing IS literature was undertaken. The pertinent literature is used as a foundation to outline the topic agenda used in the focus group sessions, while refrained from the actual discussion.
- *Stage 2*: A focus group session was conducted. The focus group discussion was held, involving business executives from IS functional areas of which the executives are expected to use the internet-based tool frequently in their daily tasks.
- *Stage 3*: Data collected (in handwritten notes and audio tapes) from the focus group session was firstly compiled with the assistance of the computer. The results were thereby analysed in detail. The adoption of thematic categorisation (Blackburn and Stokes, 2000; Boyatzis, 1998; Stewart and Shamdasani, 2000) involves the articulation and construction of potential themes and facets of constructs that

correspond with the research questions. The potential themes were considered as a reflection of the attributes or dimensions.

5 Focus group study results

The focus group session was held with six IT practitioners in Hong Kong. A large amount of ideas and opinions were sought in the study. In essence, the practitioners are satisfied with the existing tools, in regard to its low cost (usually free of charge) and ease of use. However, the information quality in terms of amount, currency and presentation has presented limitations to their business decisions. The participants reveal that Redips can add value to their managerial tasks, as it streamlines the information search, and enables effective assessment of the web communities and stakeholders. A common assertion was sought that they will use Redips in addition to the existing search practice.

The practitioners had named Yahoo and Google as two commonly used internet-based tools in their daily work. There are some extraordinary occasions that they may seek information from other search tools (such as www.altavista.com) given the two common tools did not provide adequate information for their purposes.

As mentioned before, one of the advantages of focus group discussion is to provide analytical generalisation of what participants think and perceive (Law et al., 2002; Morgan, 1996), therefore, we assimilated nine effective and desirable system attributes from the opinions sought in the focus group discussion (Table 1). Redips' characteristics generally shows a good match with the desired attributes expected by the IS practitioners. The attributes are ranked according with the number of respondents who solicit the respective aspect. Therefore, a higher number in the bracket indicates a higher consensus being conceived in the discussion session. For example, all participants found that Redips provided efficient and prompt response in the search process. A lower number reveals that some participants may have neutral perceptions or different ideas towards the respective attribute. For instance, a lot of participants found that their technical knowledge is rich and extensive, thus they did not perceive the help function embedded in Redips is highly significant.

Contrast to our result-oriented system design, the IS practitioners who are experienced in system development and engage in business planning share similar perception and valuation criteria that are close to those of end-users'. Most of the participants showed empathy on the business managers' pressure in time management. A system analysis manager stated that the tool should be time saving and with results displayed with highly relevant, clear, easily be comprehended and useful content. Two other participants supported his views as they claimed the business development managers in their organisations are 'extremely' hard working as they usually work for more than 54 hours per week, thus they cannot afford information search process that 'wastes' their time. Therefore, they stated that the search process should be time-effective, with minimum search and loading time. They commented further that long loading time will increase the tendency to abort the search process, thus, unfavourably affect managers' time management.

Table 1 Desirable attributes of an internet-based search tool and the characteristics of Redips

Desirable attributes	*Details and descriptions*	*Redips*
Time (6)	Short and prompt search time	Quick response and fetching time
Cost (6)	Free of charge	Free of charge
Search results (6)	Links should be ranked with relevance and currency of the information	Web pages are fetched in real-time to avoid outdated indexes
Learning efforts (5)	Minimum, ease of use without attending formal training	Easy to use, logical and order search process
Technical resources (4)	No additional software should be installed	Requires installation
Interface (4)	User-friendly and attractive	Practical and clean, user-friendly buttons
Language (3)	Multiple languages	Limited to English
Personalisation (2)	Customisation of specific search requirements, prior search records	Three search engines (including Google, MSN and AltaVista) can be previewed at the same time
		Phrase selection to remove unrelated pages
Assistance (1)	Help, search tips and data recovery	Help menu available

While discussing the approaches in information search or business intelligence, the practitioners stated that their search approaches vary from task to task. One IS department assistant manager stated that if he looks for new business opportunities that he has little experience about or a request is made on ad hoc basis, he may seek information exhaustively until the relevance of the search results diminishes. He explained,

> "I will seek help from experienced staff within the organisation. Yet, sometimes good business ideas are not generated from internal people as they are indulged to the existing success. Those competitors, discussion groups provide lots of comments to our existing services, despite their words are unpleasant. Listening to our customers can let us know our inefficiency and what we need to do more."

Otherwise, if he handles regular review of market positioning, he will approach a satisfactory manner. He stated "my personal experience is rich and validated in my past success. Therefore, I know when and where to get the relevant information for routine tasks, and know when to stop".

In regard to the propagation of Redips, a senior system engineer pointed out that a good design does not make the system successful and well accepted by users. He stated, "Our colleagues have a strong social network which shared lots of latest information and users' feedback on various tools". Therefore, he suggested

> "the success of Redips is not a mere concern of physical design and functionality; rather it is determined by the consumers who foster a good perception at the beginning. If the comment is positive, the word-of-mouth will leverage users' interest and attention to Redips."

A senior software development manager put forth the opinion and stated that management should not underestimate the power of management commitment and network effect. She gave an example that a new software system was well accepted in their organisation, despite the brand equity of that software system is not high. It is because the management has shown a strong support, clear vision and commitment to deploy across all departments. Middle management also plays an important role to monitor the roll out of the system. Furthermore, she recalled "most of our colleagues believe that they should get on the same board (i.e., adopt the new system) in order to become one of the majority".

6 Conclusion

The internet-based backlink search tool Redips provides an alternative way to assess various web communities and infer business intelligence to managers. In order to enhance its practicability and usefulness to business managers, the current study, as an exploratory work adopted focus group method to drill in various perceptions and views from a group of IS practitioners. In essence, they find Redips useful and assert its value in business planning and decision making. Our present study will be continued and extended with more focus group discussions composed of diverse professional participants or users. The nine attributes and specific findings about managerial problem solving approaches will also be incorporated such that Redips can be widely adopted.

Acknowledgements

This project has been supported in part by the following grants:

- HKU Seed Funding for Basic Research, "Searching the World Wide Web Backwards for Business Intelligence Analysis", January 2005–December 2006.
- HKU Seed Funding for Basic Research, "Using Content and Link Analysis in Developing Domain-specific Web Search Engines: A Machine Learning Approach", February 2004–January 2006.
- NSF Digital Library Initiative-2, "High-performance Digital Library Systems: From Information Retrieval to Knowledge Management", April 1999–March 2002.

We would like to thank the IS practitioners for their participation in our experiment.

References

Blackburn, R. and Stokes, D. (2000) 'Breaking down the barriers: using focus groups to research small and medium-sized enterprises', *International Small Business Journal*, Vol. 19, No. 1, pp.44–67.

Boyatzis, R.E. (1998) *Transforming Qualitative Information: Thematic Analysis and Code Development*, Sage, Thousand Oaks, CA.

Bradford, R.W., Duncan, P.J. and Tarcy, B. (1999) *Simplified Strategic Planning: A No-Nonsense Guide for Busy People Who Want Results Fast!*, Chandler House Press. AUTHOR PLEASE PROVIDE LOCATION.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia.

Caldeira, M.M. and Ward, J.M. (2002) 'Understanding the successful adoption and use of IS/IT in SMEs: an explanation from portuguese manufacturing industries', *Information System Journal*, Vol. 12, No. 2, pp.121–152.

Chau, M. and Chen, H. (2003) 'Comparison of three vertical search spiders', *IEEE Computer*, Vol. 36, No. 5, pp.56–62.

Chau, M., Shiu, B., Chan, I. and Chen, H. (2005) 'Automated identification of web communities for business intelligence analysis', *Proceedings of the Fourth Workshop on E-Business (WEB)*, Las Vegas, Nevada, USA.

Chen, H., Fan, H., Chau, M. and Zeng, D. (2002) 'MetaSpider: Meta-searching and categorization on the web', *Journal of American Society for Information Science and Technology*, Vol. 52, No. 13, pp.1134–1147. AUTHOR PLEASE CHECK WHETHER THE YEAR OF PUBLICATION IS 2001 OR 2002.

Chen, H., Schufels, C. and Orwig, R. (1996) 'Internet categorization and search: a self-organizing approach', *Journal of Visual Communication and Image Representation*, Vol. 7, No. 1, pp.88–102.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From data mining to knowledge discovery: an overview', in Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.): *Advances in Knowledge Discovery and Data Mining*, MIT Press, Cambridge, Mass, pp.1–36.

Feldman, R. and Dagan, I. (1995) 'Knowledge Discovery in Textual databases (KDT)', *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, August 20–21, AAAI Press, pp.112–117.

Google (2005) *Google Web Search Features*, Retrieved January 6, from http://www.google.com/help/features.html.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers. AUTHOR PLEASE PROVIDE LOCATION.

Hearst, M.A. (1997) 'Text data mining: issues, techniques, and the relationship to information access', *Presentation Notes for UW/MS Workshop on Data Mining*. AUTHOR PLEASE PROVIDE LOCATION.

Johnson, R.J. (1994) *A Cognitive Approach to the Representation of Managerial Competitive Intelligence Knowledge*, Doctoral Dissertation, The University of Arizona. AUTHOR PLEASE PROVIDE LOCATION.

Kohonen, T. (1995) *Self-Organizing Maps*, Springer-Verlag, Berlin.

Krueger, R.A. (1998) *Developing Questions for Focus Groups: Focus Group Kit 3*, Sage. AUTHOR PLEASE PROVIDE LOCATION.

Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1998) 'Trawling the web for emerging cyber-communities', *Proceedings of the Eight International World Wide Web Conference*. AUTHOR PLEASE PROVIDE PAGE RANGE.

Law, Y.F.D., Lee-Partridge, J.E., Beng, C.H. and Fen, W.M. (2002) 'Exploring knowledge management perceptions of human resource and business managers in Singapore', *Journal of Information and Knowledge Management*, Vol. 1, No. 1, pp.79–90.

Morgan, D.L. (1988) *Focus Groups as Qualitative Research: Qualitative Research Methods Series*, Vol. 16, Sage. AUTHOR PLEASE PROVIDE LOCATION.

Morgan, D.L. (1996) 'Focus group', *Annual Review of Sociology*, Vol. 22, pp.129–152.

Prescott, J.E. and Smith, D.C. (1991) 'SCIP: who we are, what we do', *Competitive Intelligence Review*. AUTHOR PLEASE PROVIDE FULL DETAILS.

Reid, E.O.F. (2003) 'Identifying a company's non-customer online communities: a proto-typology', *Proceedings of the Hawaii International Conference on System Sciences*, January 6–9, Big Island, Hawaii.

Salton, G. (1986) 'Another look at automatic text-retrieval systems', *Communications of the ACM*, Vol. 29, No. 7, pp.648–656.

Schermerhorn, J.R. (2001) *Management*, John Wiley & Sons, Inc. AUTHOR PLEASE PROVIDE LOCATION.

Selberg, E. and Etzioni, O. (1997) 'The MetaCrawler architecture for resource aggregation on the web', *IEEE Expert*, Vol. 12, No. 1, pp.11–14.

Stewart, D.W. and Shamdasani, P.N. (2000) *Focus Groups: Theory and Practice*, Sage, California.

Sutcliffe, A.G., Ennis, M. and Hu, J. (2000) 'Evaluating the effectiveness of visual user interfaces for information retrieval', *International Journal of Human-Computer Studies*, Vol. 53, No. 5, pp.741–763.

Tan, A.H. (1999) 'Text mining: the state of the art and the challenges', *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*. AUTHOR PLEASE PROVIDE LOCATION.

Tolle, K. and Chen, H. (2000) 'Comparing noun phrasing techniques for use with medical digital library tools', *Journal of the American Society for Information Science*, Vol. 51, No. 4, pp.352–370.