Leading Edge Forum
Research Grant 2003/4

Web Mining

Between 80 and 90% of the information that we work with on a day-to-day basis is unstructured and textual. Much of this information is now becoming accessible via web based technologies. Web Mining is a technology, application and product suite that can enable us to discover hidden wealth locked within unstructured repositories. Web mining promises to provide us with new levels of business intelligence, smarter alerts and new product designs. The opportunities are limitless.

Laurence Lock Lee
Australia Group

TABLE OF CONTENTS

INTRODUCTION

The World-Wide Web is the world's largest information repository, accessed by some 700million people world wide with estimates of a world wide audience reaching 1 billion within the next 12 months[1]. Barely a decade has passed since the Internet became available for public use. Its overwhelming impact on business and society at large would comfortably place it within the top echelon of technology disruptions throughout recorded history. The rapid growth of the Internet could be largely attributed to the loose governance structure, which beyond some control over domain names, is open to be freely added to by anyone with access to a computer and an Internet connection. Central to harnessing the value and gaining benefit from the Internet is being able to rapidly find or derive information that you can leverage for your benefit, whether it be scientific research, retail shopping, medical diagnosis or even on-line socialisation.

Unlike the managed data repositories that have been associated with computers since the 1960s, the World-Wide Web is only lightly governed. The majority of data is textual and in natural language form. The immense growth rate is providing both challenges and opportunities to assist both individuals and businesses to profit from its use. Web Mining (WM) is the term coined to describe the methods and tools used to both retrieve and discover information contained within the Web. Drawing its heritage from Information Management & Retrieval and Artificial Intelligence research, Web Mining is being promoted as the technology or tool of choice for leveraging the immense opportunities afforded by the Internet, as its growth continues untethered. The ratio of unstructured to structured/managed information repositories is accelerating to the extent that most of the information we use in a day-to-day sense comes from unstructured sources like the Internet or our own Intranets and private storage spaces.

A random survey of over 550 CSC staff has shown that as much as 87% of information used on a day-to-day basis is drawn from unstructured information sources. Close to 50% of information was stored in private spaces and around 89% in largely unstructured spaces. The opportunity for business improvement clearly lies in how we work with largely unstructured information spaces. Web mining has an important role to play.

In this research report, section 1 will provide a review of the "state-of-the-art" in WM, both from a technical and applications point of view. The report will also comment on product trends i.e. what developments we are seeing in the marketplace. For this report, WM is seen as applicable to any web based repository, whether it be available through the Internet or an Intranet. Section 2 will report on the results of some case study experiments conducted as part of this research or identified within the literature. These case studies are aimed at illustrating the breadth of applications available for WM. Section 3 provides a discussion on the trends and the opportunities that WM might provide.

[1] http://www.glreach.com/globstats/

1 LITERATURE REVIEW

Web Mining as a discipline is still considered to be in its infancy. Much of the foundation research supporting WM has been drawn from Information Management & Retrieval (IM&R) and Artificial Intelligence (AI). From IR research on taxonomies, thesauri, classification systems and from AI, research on knowledge representation schemes like semantic nets, production rules, frames and scripts, intelligent agents, machine learning and data mining are all contributing to the WM field. WM also inherits many of the lead application areas from data mining, such as fraud detection, customer behaviour analysis and Customer Relationship Management (CRM). Visualisation of "mined" artefacts is a complementary field, where the challenge is to provide mechanisms for the end-user to be able to interpret the WM results (Yang et al, 2003). The following taxonomy of Web Mining illustrates how the field of application is broadly broken up into web content mining and web usage mining (Cooley et al, 1997):

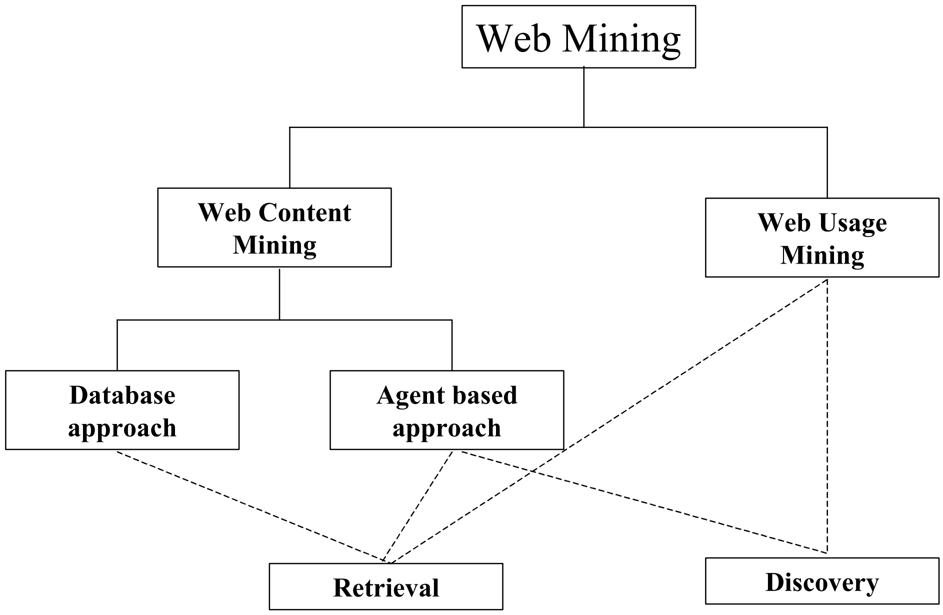

Figure 1 – Web Mining Taxonomy

Web content mining typically involves searching through web based information repositories. Broadly speaking, there have been two approaches to searching content. The data base approach is an extension of traditional data base querying, though adapted for the nature of the Web (OKS Group, 2003; Biggs, 2003). The agent based approach is inspired by autonomous agent research and the creation of web-bots for adaptively searching or spidering through web repositories (Menczer, 2003). The driving force behind Web usage mining was business to consumer (B2C) e-commerce. By searching for usage patterns from within weblogs, one could adapt a web site to maximise the opportunity for increased sales (Doherty,2000; Iyer et al, 2002; Manchester, 2001; Mulvenna et al, 2000; Thelwall, 2001). Both web content and web usage mining can take a retrieval approach of known or anticipated content or a discovery of new or unexpected content. There is some debate over whether content retrieval is true web mining (Hearst, 1999), however on balance, the term has been typically used to cover both aspects.

The Special Edition on Web Retrieval and Mining for Decision Support Systems (Chen, 2003) identifies a strong future for WM, with the prospect of applications supporting a multi-lingual web, a multi-media web and a wireless accessible web. The following sections will review the literature in terms of technology trends, application trends and product trends.

1.1 Web Mining and Data Mining

Web mining draws its genesis from data mining. This section reviews the basic principles of data mining and attempts to characterise how web mining builds on these foundations. Figure 2 – Web and Data Mining, provides a simplified overview of a "knowledge discovery" process within which data mining is a step in the process. Knowledge Discovery could be defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al, 1996). Traditional knowledge discovery has been conducted from data extracted from large data bases typically holding transaction data for applications like target marketing, fraud detection, fault diagnosis etc.. These data sources are clearly more structured than the web and the contents constrained to either numeric or simple text labels. Prior to data mining the data needs to go through processes to check for erroneous data and to transform the data to a form that can be processed by the different data mining engines. Data mining is substantially a statistical modelling process. The outputs can take many forms, from tables of figures to graphs, plots or multi-dimensional images e.g. contour maps. Typically the data mining exercise is aimed at classifying a particular pattern or feature with a categorisation of interest e.g. potential loan defaulter, potential fraud, likely high value customer, a particular fault in a machine etc.. An alternative data mining output can be in the form of decision rules or trees induced from the data. Unlike the statistical processes, rule induction algorithms provide output that describes the hierarchical path of inference taken to achieve a prediction result.

One of the key differences for web mining is the less structured data source. The structure of the web is far from a neat set of tables with clearly defined attributes of known type. The main structural element for the web is the URL. The content is also more textual than numeric, which does not lend itself as well to algorithmic analyses.

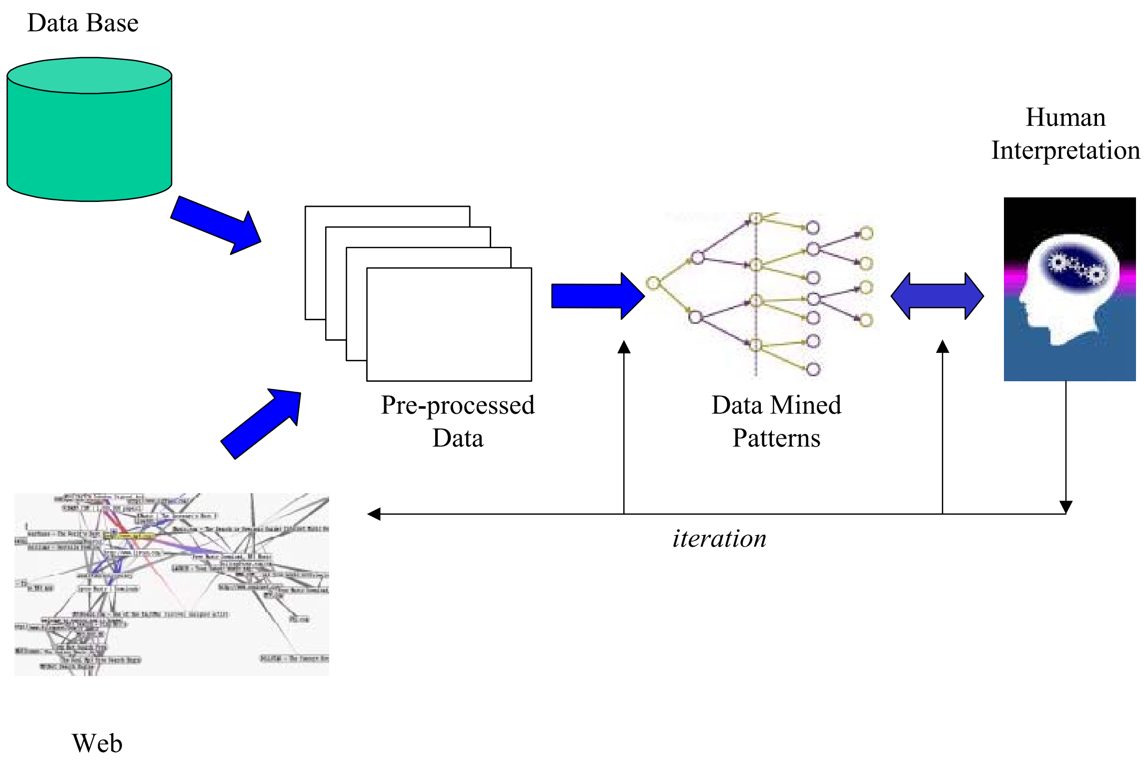

Figure 2 – Web and Data Mining

However, the trend towards web enabling all applications and information sources will eventually see data mining sources become just another entity on the web. The earliest applications of data mining to the web are to mine for “clickstream” patterns. In this case each “click” is considered a transaction from a finite set of possible transactions on a particular e-commerce web site. This data can be pre-processed into a traditional database and data mined with traditional data mining tools to identify high potential customers. A variant of the clickstream analysis is the affinity analysis, where understanding what particular customers buy in common, can identify good cross selling opportunities. The Amazon.com “people who have bought this book have also bought…..” is an example of mining web transactions.

Of more recent interest is web content analysis. The web search and metasearch engines are now capable of clustering or classifying content with either linguistic or linguistic pattern matching algorithms. These classification engines are labelled "text miners" and perform an analogous process to data mining engines, but with textual data sources. The outputs are typically clusters of concepts or topics arranged in a hierarchy of folders containing URLs. SAS provides a good example of how text mining can be combined with conventional statistical processes in an application which identifies the impact of physician practices and patient outcomes that are largely documented in text. (www.sas.com/news/feature/22dec03/uol.html). In this case the physician practices would be a classified set of classes, as would the patient outcomes. The classes of physician practices and patient outcomes could then be treated using traditional correlation or regression analyses.

In summary, web mining applications are typically built on the existing foundations of data mining, making use of many of the same functions. However, the pervasiveness of the source data and the emerging ability to classify textual sources opens the door for many more innovative and novel applications.

1.2 Technology Trends

Web Mining technology draws much of its heritage from the fields of Artificial Intelligence and Information Management & Retrieval. It could be argued that these fields meet around the area of text mining, which given the textual nature of the web, is the core technology underpinning web mining developments.

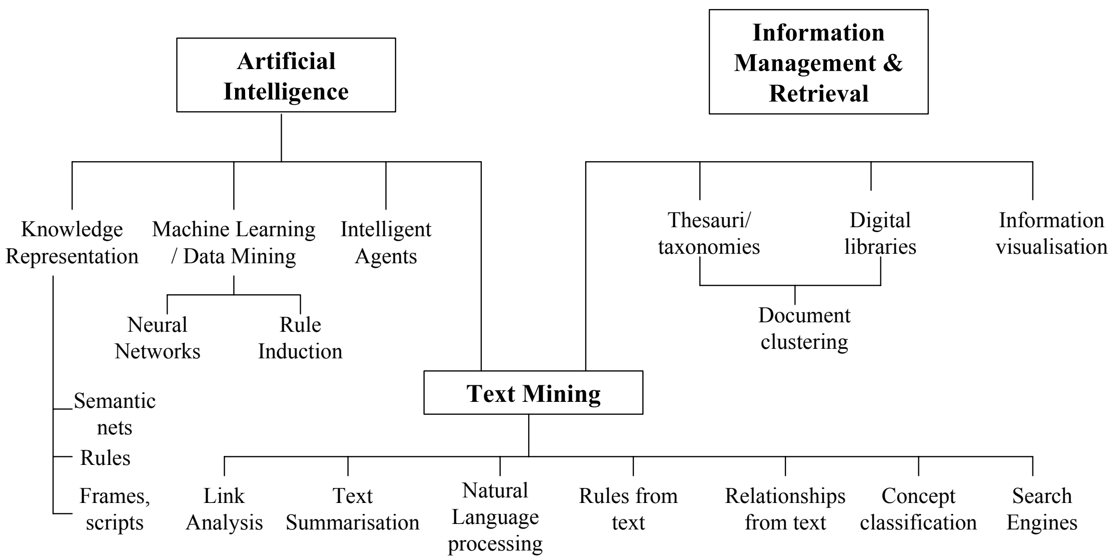

Figure 3 – Web Mining Technology Taxonomy

The above figure attempts to classify the contributing technical areas supporting web mining, though it is noted that many of these areas overlap. Looking closely at the research literature specifically supporting WM, the issues of link analysis verses textual analysis has been highlighted, especially in the context of search engine effectiveness (Thelwall, 2001; Getoor, 2002; Menczer, 2003). Concept classification has proved a popular area for search engine vendors, enabling concepts to be identified where potentially a textual match has not been made (Chuang and Chien, 2003; Roussinov and Zhao, 2003). Clustering into generic concepts has also been specialised to identify communities of end users and/or relationships between those who may share a common interest (Thelwall, 2003; Gibson et al, 1998; Rousisinov and Zhao, 2003). Alternatively, clusters could be expanded into full taxonomies or knowledge maps, providing a structured overlay for the web (Chuang and Chien, 2003; Wang, 2002).

Research on natural language processing in the context of WM is aimed at enabling end users to make natural language queries of the web. For more structured queries the development of WebQL, a web based extension of SQL for data bases is being developed (Biggs, 2003; QL2Software, 2003). The alternate approach is to match concepts mined from the query itself with those in the repository. In this situation a more verbose query, rather than a restricted query is encouraged, to provide the search engine with sufficient text to characterise concepts for matching (Butler, 2003).

Once a search has been conducted one still has the issue of how best to present the results of the search to the end user. Most search engines simply provide a list of URLs from which the end user can typically view only a small percentage. Graphical tools are now being developed to provide the end user with a contextual map of their search results, with an ability to drill further into areas of stronger interest (Yang et al, 2003; Adar and Tyler, 2000). The use of WebQL is also being promoted for its ability to format search results more usefully e.g. in tables or charts (QL2Software, 2003).

The web is increasingly becoming a source of visual as well as textual data. The ability to search for specific images residing on the web has attracted research attention. While typically the search still relies heavily on context provided by surrounding text, some level of image processing is also now being incorporated into image based searches (Chen et al, 2001; Kou et al, 2003).

A final area of note is the web mining of interactive collaboration spaces. These spaces might be discussion forum areas, electronic meetings and even e-mail logs (Getoor,2002; Roussinov and Zhao, 2003; Bontis et al, 2003; Lock Lee, 2003). The objectives of relationship mining can range from identifying patterns of social networks i.e. via say, community membership, through to identifying opinions on consumer products in chat rooms.

While web mining research and technologies builds on a solid base of AI and IM&IR research there are still many challenges and lessons to be learnt. At this time web mining, like data mining requires a level of specialised expertise not normally found within the user community. Much of the data collected in weblogs still goes unanalysed, with the challenge of integrating data sources with business interpretation remaining (Kavovi et al, 2004). The relative immaturity of WM research however, has not deterred both established and new product vendors incorporating new features based on these developments. This aspect will be explored further in section 1.4.

1.3 Application Trends

The most dominant application area for WM is related to Internet based e-commerce (business-to-consumer) and web based customer relationship management (CRM). The e-commerce boom of the late 1990's and the plethora of Internet business start-ups provided the fuel to launch web usage mining as the dominant web mining application. WM provided the facility to track customer behaviour for web based businesses more comprehensively that in any previous business model. The ability to rapidly adapt web sites, product information and even pricing was available to those who could both collect and rapidly analyse consumer navigation patterns (Iyer et al., 2002). Web site navigation could be personalised to provide a unique experience for the consumer (Mulvenna et al,2000). Consumer forums could be established to enable them to provide direct feedback on products and/or suppliers.

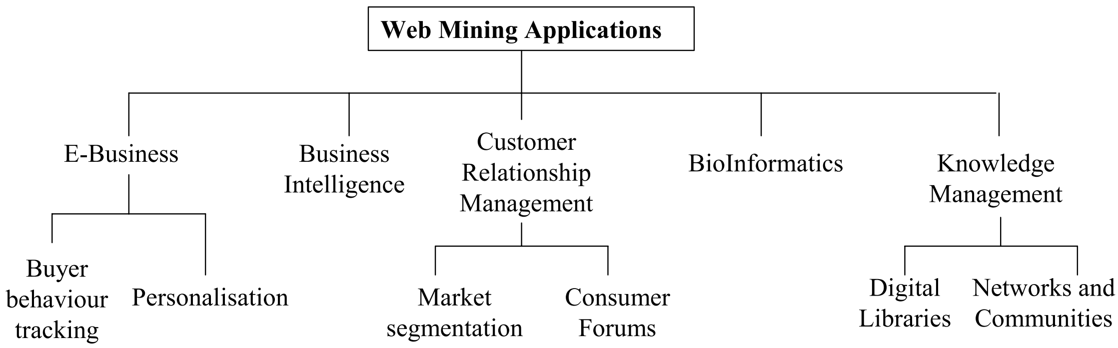

Figure 4 – Web Mining Applications

Research has been conducted into applications that make use of textual information on the Web, to assist in predicting stock movements (Fung et al, 2003; Wuthrich et al, 1998). By reviewing the plethora of financial news sources, researchers were able to associate common phrases with particular market movements. Weightings are derived for particular combinations and inference rules can be built around training data sets. Researchers were able to simulate predictions based on the combination of numerical and textual input, that could clearly outperform the market.

Applications in the field of knowledge management could be categorised as "connections" or "collections". Identifying common interest communities can occur indirectly e.g. the way Amazon infers a common interest by identifying those that have bought the same book and therefore recommending to others, books that the interest community has also bought. Membership of forums or news groups is a more direct method of identifying a common interest community. The web could be considered one vast digital library. The function of a librarian is to make it easier for people to find information. WM could be viewed as the digital library's librarian, or at least a powerful tool to those acting as their own librarian.

As well as becoming a vast repository for documents, the web is increasingly becoming an electronic platform for social interaction or "connections", for business or pleasure. Chat rooms, news groups, discussion groups, mailing lists, e-mail, instant messaging etc. provide both the mechanism for human interaction on the web and usually a data trail that can potentially be mined. The rapid adoption of weblogs (personal web based diaries), that are typically rich in weblinks, has caused problems for the Google search engine by swamping its hits rankings. Applications are now emerging for specifically searching weblogs (http://www.feedster.com/). Visualisation applications are also now emerging which will enable users to visualise the conversation space that they are participating in. An example application called "social circles" is able to dynamically build a visualisation of social circles from your mailing list subscription:

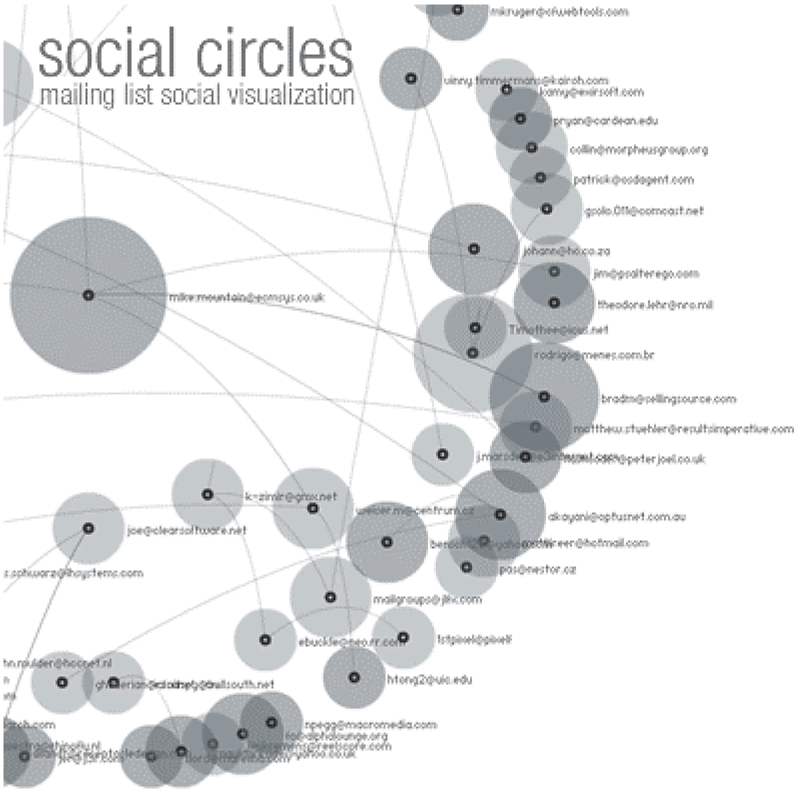

Figure 5 – Social Circles (www.marumushi.com/apps/socialcircles/)

Business Intelligence systems have traditionally relied on internally developed repositories. With the advent of the Web and WM, the opportunity exists to extend the Business Intelligence repository external to the organisation. For example, systems have been built to analyse Securities and Exchange Commission electronic filings, looking for potential items of interest to investors or business partnerships (Gerdes, 2003). The larger information base afforded by the web, while less structured, provides the opportunity to become pro-active by using WM tools to deliver business intelligence before it is requested i.e. push technology. The integration of WM and traditional data warehouse based business intelligence systems therefore becomes critical to achieving user acceptance (Kanzler, 2002).

The final application area identified has been labelled “Bioinformatics”. The extensive literature available in the biotechnology field has made it difficult for researchers and drug designers to comprehensively cover, when exploring for new drug development opportunities. Identifying protein patterns, exploring gene structures and their interaction are just a few of the applications making use of WM techniques to complement traditional data based searches (Hanish et al., 2003; Kou et al., 2003, Hearst, 1999). The growth in medical web portals is also providing opportunities for WM. Chen reports on the “HelpfulMed” system which provides a single search interface to multiple medical web portals, resolving issues of overlapping terminologies in the process (Chen, 2002).

In summary, the trend in applications of WM is either in areas which can profit from business to consumer e-commerce or in areas where web based portals are growing rapidly and providing overlapping, and at times conflicting information. In all of these areas the end consumer is driving the demand. In the e-commerce space, Internet shoppers are driving the changes in the Web based business models. While lower prices has been the main differentiator of web based businesses to date, there is now evidence of non-price factors being taken into account by consumers e.g. security, delivery performance etc.. For areas like medical information portals, the major use is in educating the end consumer, with the intent to relieve some of the load on medical practitioner staff, by generating a more educated community.

In comparison with data mining, which traditionally works with numbers, web mining application still tend to rely on the end user for the interpretation of the data presented. For example, with a typical data mining application like fraud detection, the end user would expect the software to make suggestions on potential fraud targets. Web mining applications tend to fall short of this level of functionality. The applications might typically identify linked clusters of textual information from which the end user would identify the important patterns. SAS, however, have integrated text mining with data mining by enabling their Text Miner product to store derived category counts inside the data mining product for further processing and interpretation. (www.sas.com/technologies/analytics/datamining/textminer/) .

1.4 Product Trends

It is still early days for WM and there appears few if any products that one could say are a specialised WM product. This is not to say that there is little product support for WM, quite the contrary. WM functionality can be seen as features drawn from several other product areas. Some principal features that one might expect of a WM tool include:

- Provision of data collection, storage and management functions;
- Provision of data cleansing / filtering functions;
- Facilities for text processing and manipulation;
- Functions for identifying patterns, clusters, relationships etc.;
- Reporting facilities, both textually and graphically to communicate results; and
- An integrated development environment for iterative development of the WM solution.

The features above are often drawn from different product suites. Figure 6 – WM Product Landscape, illustrates the product suites in which one may find WM functionality. Enterprise search engine products, many of which have advanced concept search facilities, tend to fulfil the information retrieval requirements of a WM solution. Business Intelligence / Data Warehouse systems are seen as the repositories for raw material to be mined. Traditionally these data warehouses are highly structured, but business intelligence solutions are now starting to expand to incorporate unstructured and textual information typically found on the web. CRM systems could be seen as a class of data warehouse, concentrating specifically on client information. The web analytics product market is still quite specialised. While there are several small niche products becoming available, the traditional statistical analysis products like SAS and SPSS are now moving into this space.

The following section will explore the product space for each of these sectors, in the context of how they contribute to WM.

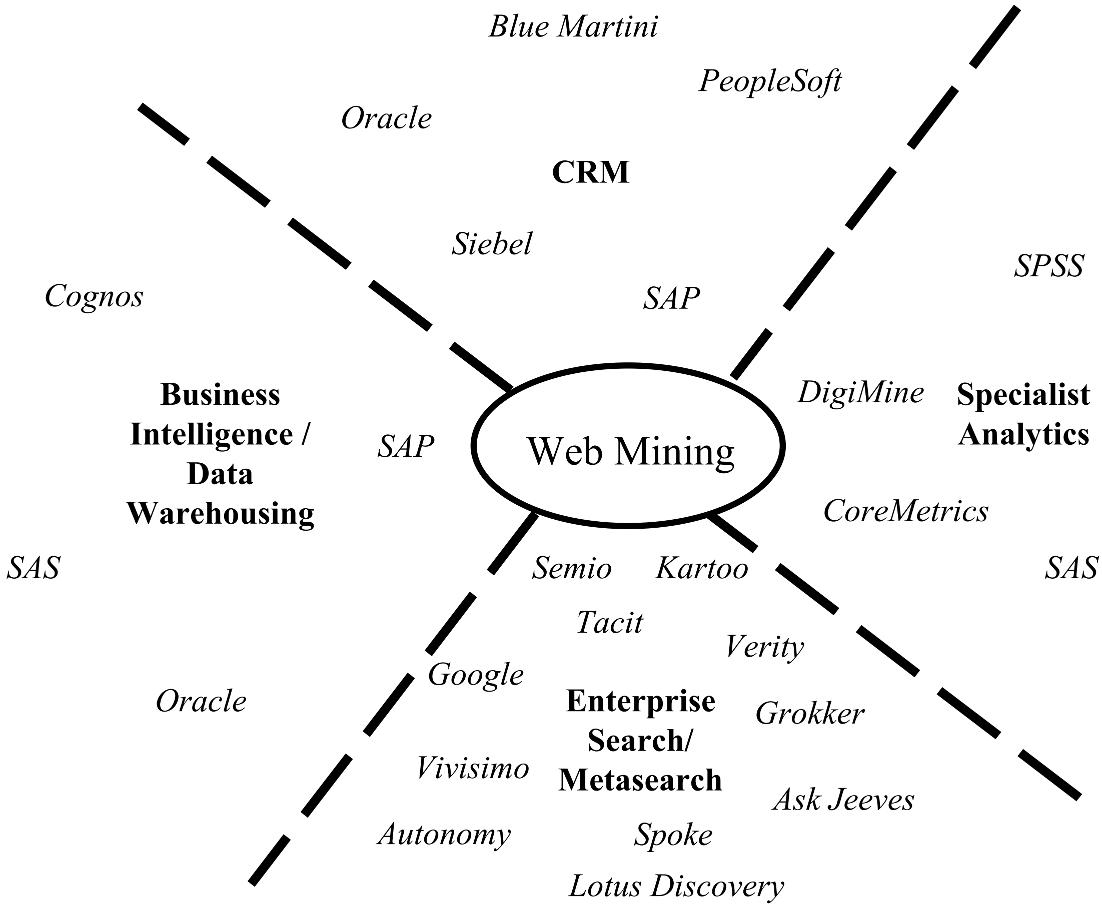

Figure 6 – WM Product Landscape

The above figure provides a sample of both new and mature products which fit within each of the application areas. Several product suites span more than a single area. Enterprise suites, like SAP, Oracle and the IBM/Lotus offerings arguably span all sectors.

1.4.1 Enterprise Search

Enterprise search is the application area that provides the core technology for web mining. Textual search engines are the core technology for web mining. The Gartner magic quadrant analysis shows a very crowded and competitive field. Text categorisation/clustering or concept identification is an important new feature for search engines. Pioneered by Autonomy in the mid 1990s, most of the mature search engines like Verity, Lotus Discovery, SAP Portal now boast a concept classification capability to different degrees of functionality.

Some emerging tools not identified in the Gartner analysis could be classified as MetaSearch Engines. Tools like Grokker (http://www.groxis.com/service/grok/), Vivisomo (www.vivisimo.com) and Kartoo (www.kartoo.com) take the results from one or several search engines and develop a categorisation of these results for display for the end user. In essence the metasearch engine provides an overall summary of the search results, rather than the typical long list of URLs provided by most search engines. Grokker and Kartoo provide the results in a graphical map form, enabling the user to explore the results space and look for important patterns.

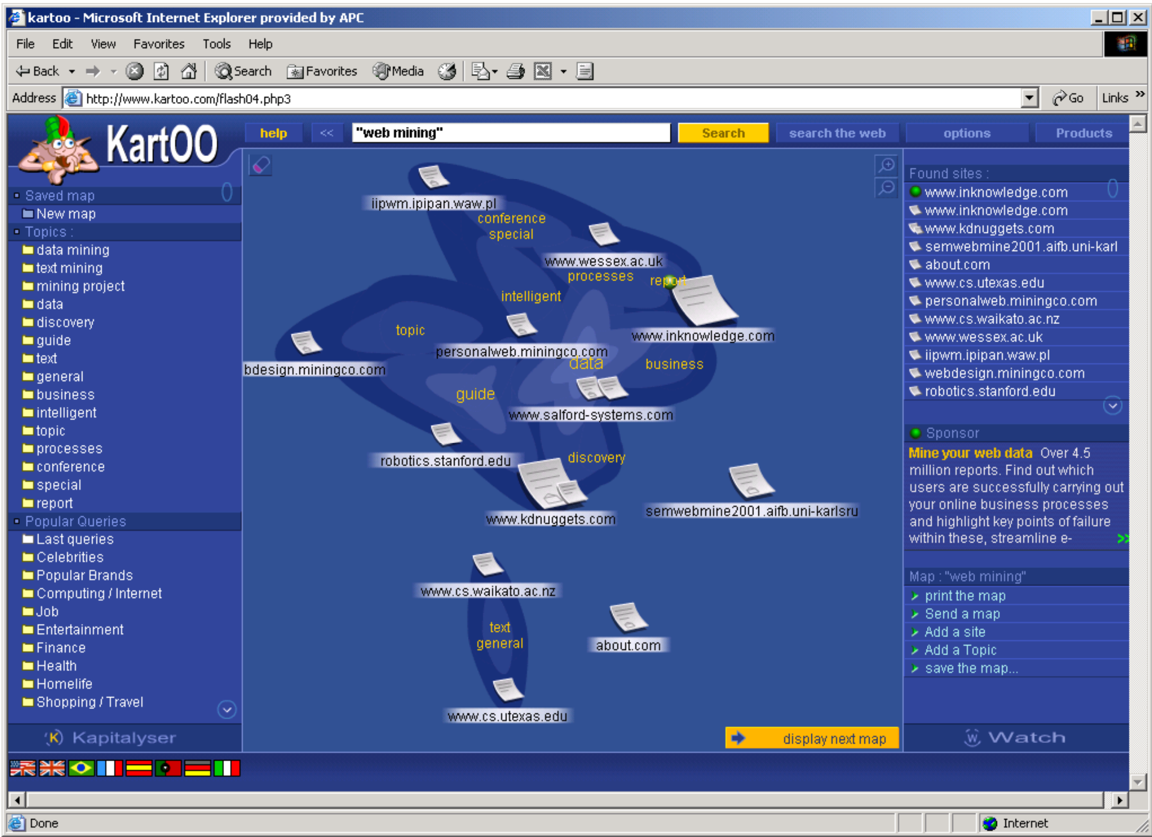

Figure 7 – Kartoo MetaSearch Results

The Kartoo results are presented in the form of a contour map. The clusters are also identified as folders on the left hand side, containing the URLs that belong to each identified clusters.

The other class of enterprise search products deal with "connections". Two examples of this class of product are Tacit (www.tacit.com) and Spoke (www.spoke.com). Both products use e-mail repositories as their source data and both are looking to identify relationships. Spoke is targeted at sales force support whereas Tacit is principally about expertise location. Both products have patented techniques for dealing with privacy issues. The other software products dealing with other forms of electronic communications identified in section 1.3 have yet to reach the commercialisation stage.

1.4.2 Business Intelligence / Data Warehousing

Some of the emerging enterprise search engine products have been identified with Business Intelligence applications. Currently, the leading products like Oracle, Cognos and SAP BW are still substantially working with numeric data. These products are all providing multiple data channels into their data warehouses, which include e-mail and the web and each provide web based user interfaces to their applications. These applications are not strictly offering web mining, but more often just reporting textual information from their data warehouses.

Enterprise software providers like SAP, Oracle and IBM/Lotus do provide advanced search capabilities through their portal products. The portal interfaces are capable of integrating the presentation of both numeric and textual information to the end user for business intelligence functionality.
Only SAS appears to be integrating their text mining functions with their data mining and business intelligence applications. As mentioned previously, the SAS text miner can be used to "code" common text based clusters and store these results in the SAS data warehouse for further processing.

1.4.3 Customer Relationship Management (CRM)

CRM could be identified as a particular class of business intelligence focussed on the customer, rather than internally. Gartner identifies Siebel as a clear leader in this field. Competition is coming from the enterprise software providers like Oracle, PeopleSoft and SAP. Siebel provides data mining functionality within its analytics software. However, there appears to be no web mining capability available. Again the enterprise software providers provide some support to web mining through their portal offerings.

New products entrants like Blue Martini (www.bluemartini.com) are focussing on web based businesses, providing a full range of CRM functionality, including analytics. Web site usage logs provide the data for the analytics functionality. Web usage data is mostly numeric and therefore lends itself well to conventional analytics and data mining.

1.4.4 Specialist Analytics

Specialist analytics and data mining product providers include the traditional statistical tool vendors like SAS and SPSS, along with the new age products like CoreMetrics (www.coremetrics.com), Doubleclick (www,doubleclick.com) and

DigiMine (www.digimine.com) who focus more specifically on web userlogs or "clickstreams". Clickstream analysis is the foundation application for web mining, providing early value for the emerging web based business market.

Once we move from weblog analysis to content analysis, then it is the textual mining capabilities that become important. Both SAS and SPSS have text mining capabilities, though SPSS appears to focus mainly on processing free form survey results. As mentioned previously, SAS integrates its text mining well with other SAS analytical functions. Despite the attraction of the web based business model, the majority of businesses today would do only a small portion of their business on the web. Hence, the web mining of clickstreams is of only peripheral interest. The generic capabilities of products like SAS and SPSS to mine both numeric and textual data on the web opens the doors to new and potentially novel applications. Some of these are explored in the next section.

1.4.5 Other Novel WM applications or products

This section identifies other emerging products that haven't been specifically covered in the previous sections.

1.4.5.1 QL2 Software (www.ql2.com)

QL2 software is the inventor of WebQL, the web equivalent of SQL. The value of WebQL is its ability to interrogate or look inside particular web pages for answers to queries in the same way as SQL is used to look inside data bases and respond to queries. The query syntax is modelled on SQL and therefore is not particularly friendly to the typical end user. However, to the experienced data base searcher, WebQL provides the facility to interrogate web pages for data or text pieces, captured into XML, spreadsheets or data bases for further processing. Like SQL, WebQL works best if it knows about the site it is searching. For example, a programmatic WebQL call could be made to extract the current stock prices for a particular company in real time, and provide an alert should the price fall outside a specified range. Embedded WebQL calls could be used to collect web based information into a data base for mining purposes. Another popular application could be to monitor a particular airline web site looking for a particular fare to be offered below a desired threshold. Biotechnology and pharmaceutical companies are using WebQL to monitor information relating to new drug development.

Perhaps the greatest potential for WebQL is in applications that require the collation of disparate information from web based repositories, whether they are internal intranet sites or from the Internet. The most common development approach is to identify the web based sites that contain the information you are looking for, interactively build the WebQL query to access, integrate and report the results you require. Once a query has been built and tested, then embed the query in an application that can be auto-scheduled to run at the required frequency. Typical end-users might be analysts or researchers who are required to sift through large amounts of information on a regular basis.

The following figure provides a simple example of selecting images and links from the CSC home page:

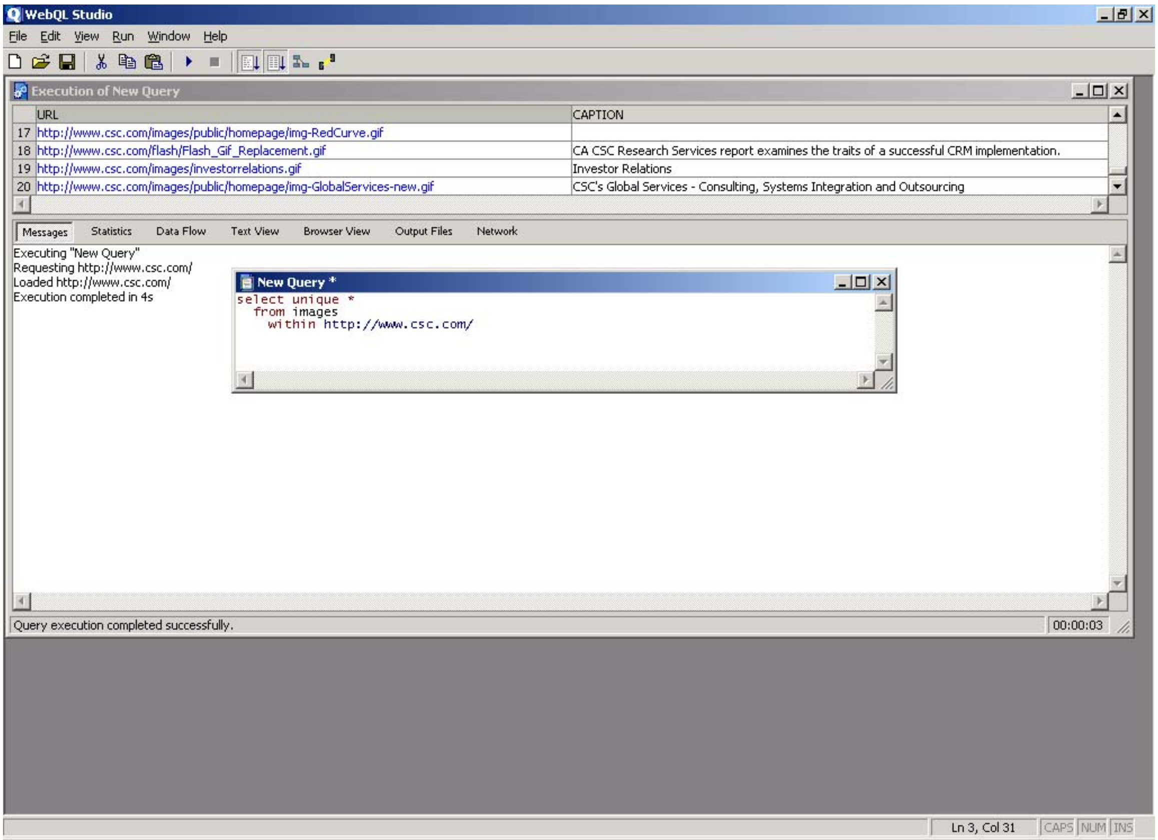

Figure 8 – WebQL of CSC Home Page

1.4.5.2 Touchgraph (www.touchgraph.com)

Touchgraph appears to be an experimental metasearch engine product which can map results from Google or Amazon. GoogleTouchgraph accepts a starting URL and then builds a map of related sites as determined by Google. The following map is the result of the www.csc.com mapping.

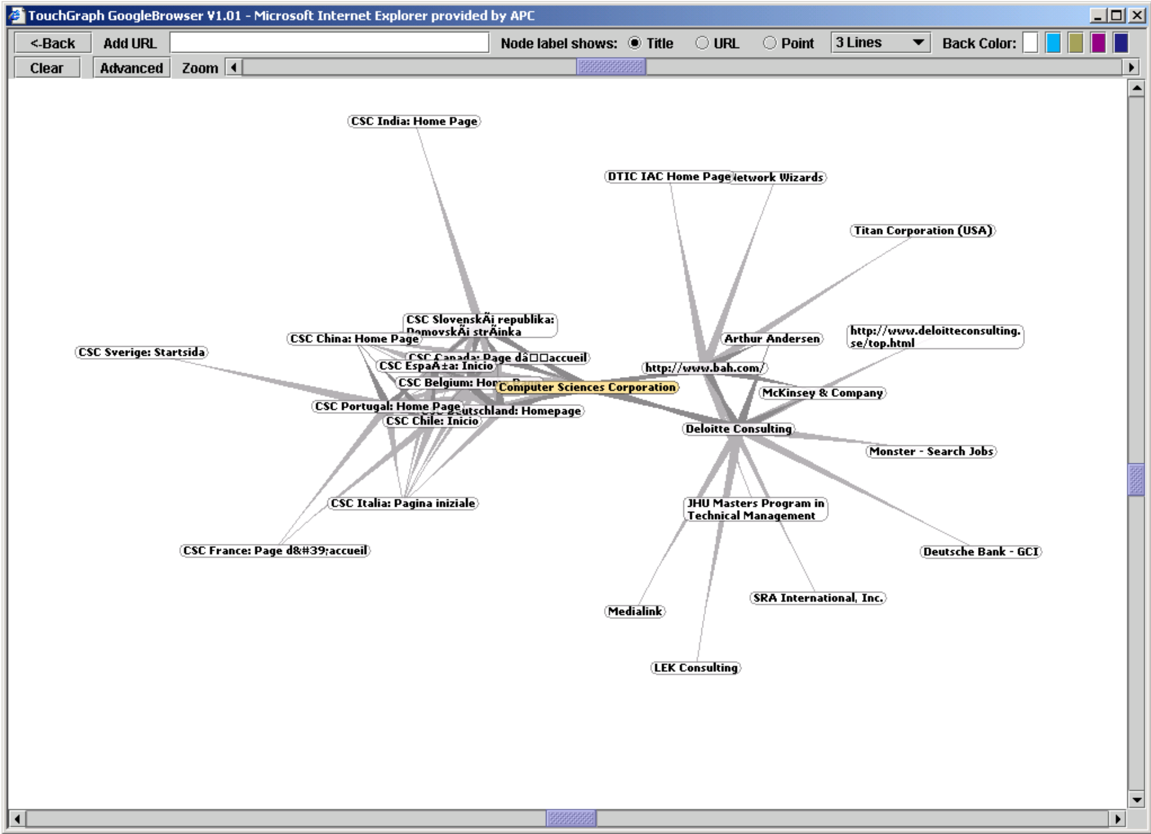

Figure 9 – csc.com Touchgraph

Other than the cluster of internal CSC sites there appears to be some strong relationships with Deloittes Consulting and Booze Allen Hamilton. As a reflection of who CSC might associate with most, it would not appear to be very accurate, as few of the hardware or software vendors are present. Looking at other sites like SAP and IBM though, the maps appear to be more representative of what one might expect. Clearly the results are impacted by the publishing policies of the individual webmasters.

The amazonTouchgraph is perhaps more useful as it can provide a full map of potential reading around a nominated topic. The following map provides a landscape of books related to web mining.

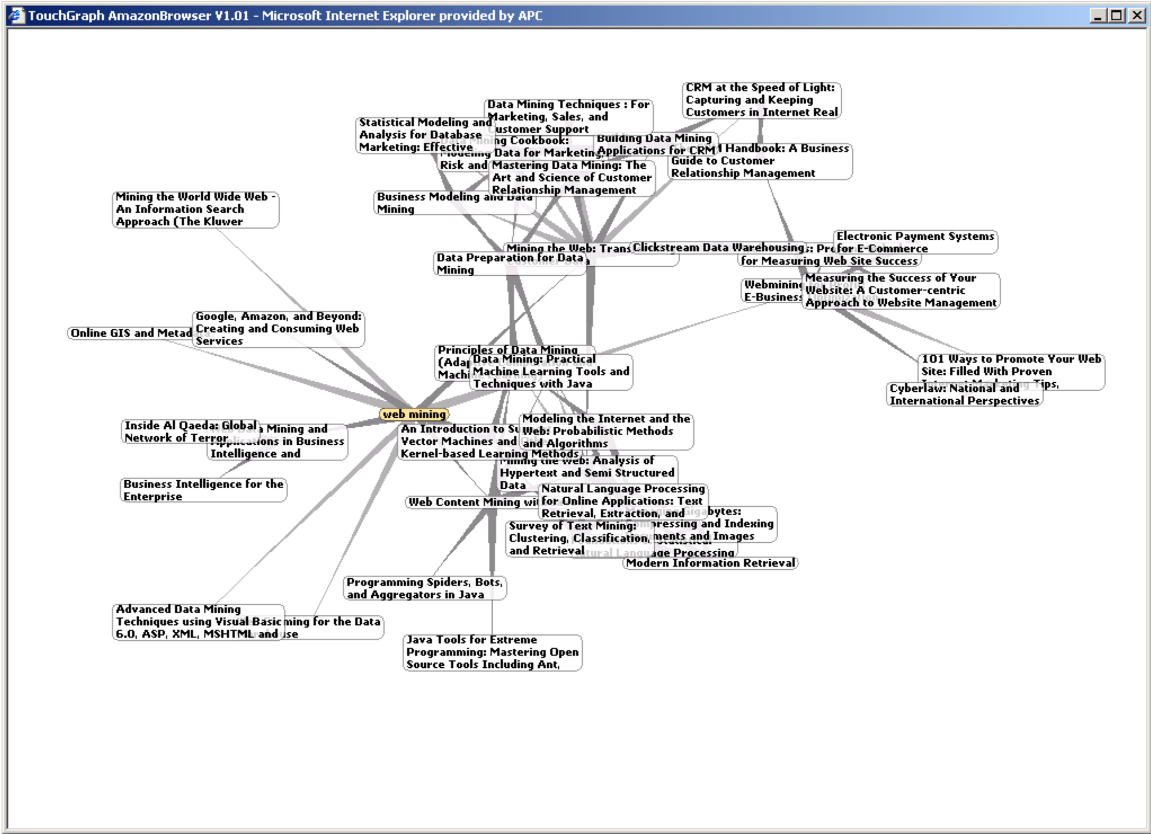

Figure 10 – AmazonTouchgraph – Web Mining

While Touchgraph appears to be still in the experimental area, one can see potential for using the software to explore networks of information, be it for business or pleasure.

1.4.5.3 Megaputer's PolyAnalyst and Web Analyst (www.megaputer.com)

Megaputer has recently released web mining products (Web Analyst and X-Sell Analyst) to support typical on-line retail sites in maximising the returns from their customer base. While these products are relatively new, they are built on the base of a fairly mature data mining package called PolyAnalyst. Polyanalyst integrates a collection of data mining and text mining techniques into a single tool. It is built on the distributed COM based architecture which enables the data mining solution to be easily integrated into larger applications or used as a stand-alone tool. The Web Analyst tool implements "web usage mining" by looking for patterns of access and navigation of a given commerce site. X-Sell Analyst focuses on identifying cross selling opportunities. Web Analyst and X-Sell largely compete with the majority of Web usage mining products on the market. Perhaps the uniqueness with Megaputer will come with the potential to marry web usage mining with web content mining offered through its PolyAnalyst offering. The text analyst component of Polyanalyst is relatively mature, in that it provides a full environment for the analyst to explore textual repositories. The text analyst algorithms perform a similar hierarchical clustering technique to other text mining products like Semio, Verity, Vivisimo etc.. However, the environment allows the analyst to conveniently explore the text space around the key concepts identified.

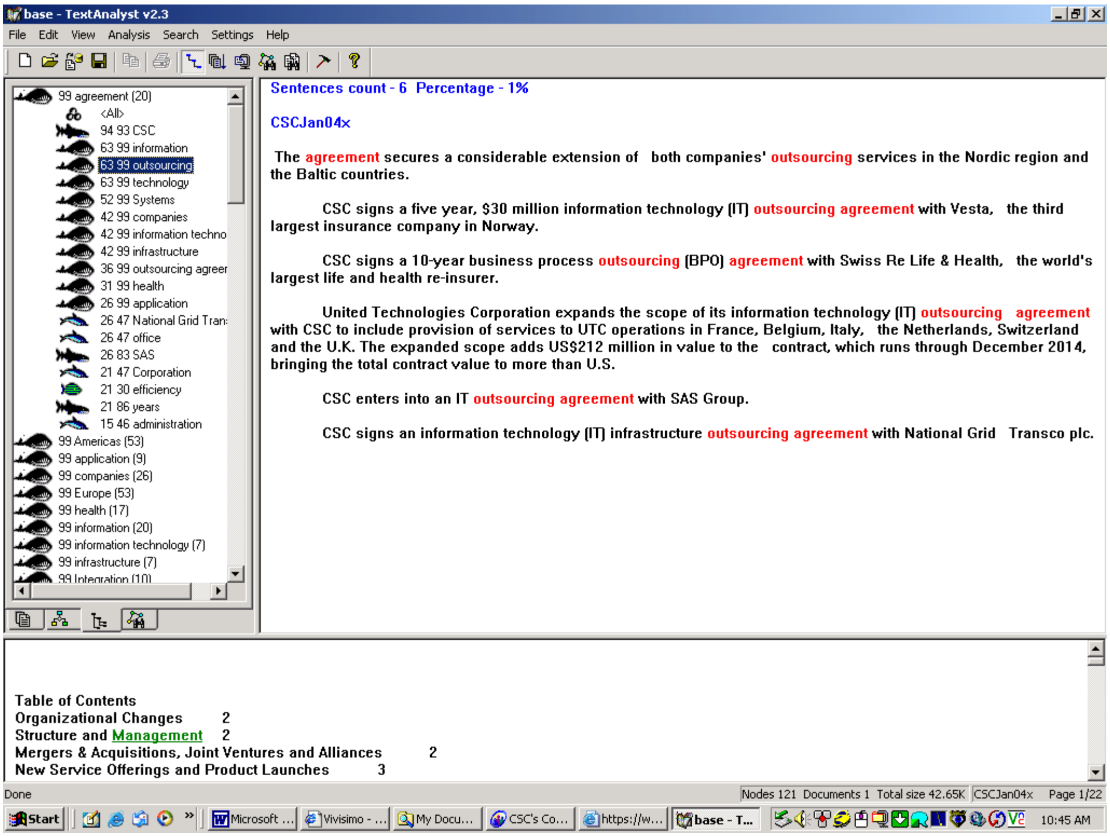

Figure 11 – Megaputer Text Analyst Environment

The above figure shows the text analyst environment. In the left hand pane one can see the hierarchy of key concepts extracted. The concept selected is "outsourcing" which is strongly related to the concept "agreement". The numbers "63 99" refer to the strength of this relationship (63) and the overall strength of the concept in the whole document (99). In the right hand pane the selection of text containing both concepts is shown. The text for this analysis was taken from a competitive intelligence report written on CSC. Similar analyses could be done on other company reports, looking for and exploring key concepts emerging from the documents. One can quickly see the power of this technology for the business intelligence and market intelligence functions.

1.4.5.4 Visible Path

Visible Path is an example of a growing number of products aimed at mining relationships and social networks. The product's focus is on large corporate environments where understanding relationship networks can enhance the sales function in particular, but could also enhance support functions. The software works by mining a company's messaging sources like e-mail, instant messaging, calendars, meeting rooms, directories etc..to develop a database of weighted relationship network information. The purported strength of the product is in the algorithms and techniques used to develop the relationship model. This information can then be used with some confidence, to navigate to others in the company that may have knowledge of, or relationships with, prospective clients. Alternatively, it could be used to identify those that may be connected to experts in a particular technical area.

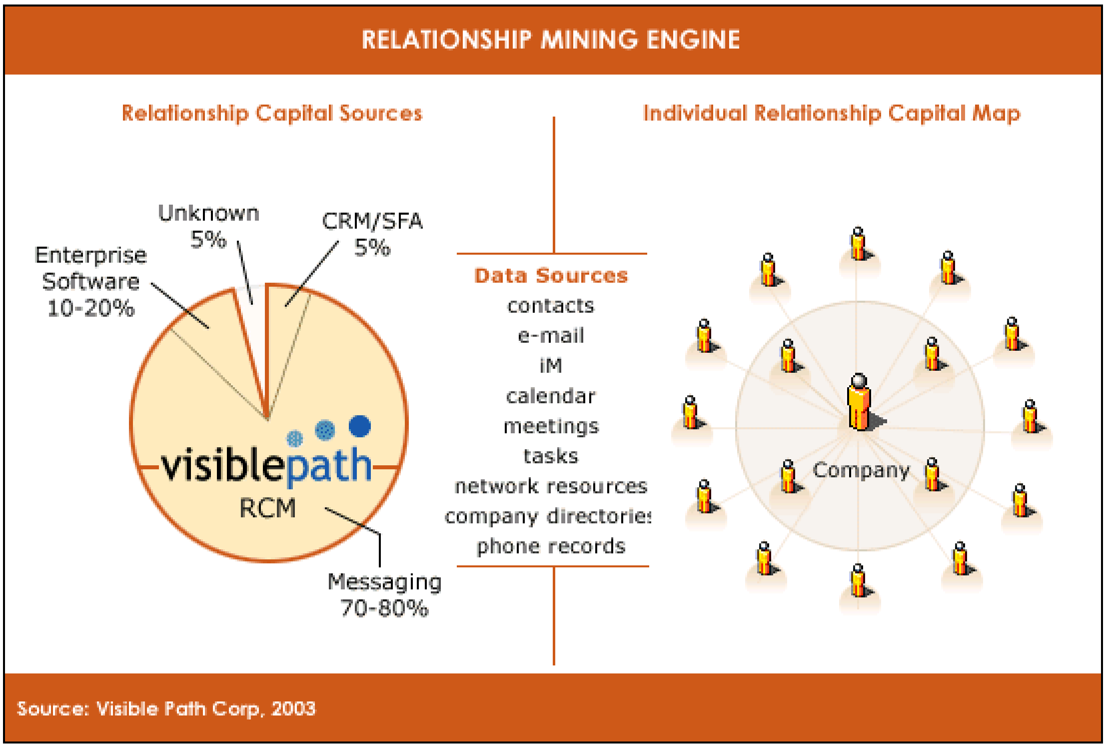

Figure 12 – Visible Path

The above figure illustrates broadly the intent of the product. Clearly privacy of information is a central concern of those considering products like Visible Path. With Visible Path and other products dealing with similar messaging sources like e-mail e.g. Tacit software (www.tacit.com), the approach has been one of individuals needing to “opt in”. In other words, individuals need to explicitly give permission for their details to be made public. With Visible Path, no relationship information is revealed without the express permission of the individuals involved, and all data is encrypted to further protect privacy.

1.4.5.5 IBM’s WebFountain (www.almaden.ibm.com/webfountain/)

The IBM Almaden research center have spent over $US100mill. developing WebFountain, a large text analytic solution which collects massive amounts of unstructured and semi-structured text, converts them to XML tagged information, prior to mining for patterns and trends. Text sources include internet data, weblogs, bulletin boards, enterprise data, licensed content, newspapers, magazines and trade journals. What is unique about WebFountain is that it stores the tagged content centrally in an IBM data facility on currently, 160 terabytes of disk space. This store then becomes a unique repository of machine-readable information that can be explored by the plethora of data mining tools that can be selected for the particular task at hand. For example, news service provider, Factiva, in partnership with IBM, are looking to provide a service for assessing company reputations based on the unstructured information gleaned from journalist articles, reports, bulletin boards and weblogs.

2 CASE STUDY REVIEWS

2.1 Mining CSC's Communities on the Portal

It is often said that a company's informal organisation is its lifeblood. Organisational structures can come and go, but long-term relationships forged around common interests have been shown to endure even the most disruptive organisational changes. CSC's communities are designed specifically to support CSC's informal organization. With over 700 communities in place and close to 20,000 members in the CSC Portal, it is now possible for individuals and groups to collaborate enterprise wide around topics as diverse as aerospace, java programming and women in technology. Current community metrics can keep track of the number of communities established, number of members and some community activity like document posts etc.. What is missing however, is information about how the communities and the individuals within them might relate to each other. How do we know which communities are truly global? Which communities are most enabling of a "one CSC"? Which individuals are likely to be playing strong knowledge brokering roles across the organisation?

This exercise is looking to characterise the informal relationship structures that might exist across CSC by mining the CSC portal and community space. To limit the size of the exercise, only community membership information was used to characterise the relationships. Therefore the nature of the relationship is simply one of a shared interest in the community subject, more so than necessarily a direct interaction. The information is fed into a social network analysis package[2] to produce a graphical view of relationships between individuals by shared community membership. The relationship information is further processed to identify those communities that are most similar, in that they share a common membership, and those individuals who appear to be best placed within the network structure to act as key brokers across the organisation, from an interest profile perspective.

The graph below provides an indication of the relationship between community membership and activity. It can be see that the top 50 most active communities appear to also be the most populous, with close to 10,000 staff being part of their membership, or roughly 50% of all staff who have registered with a community. Overall there are over 700 communities with close to 20,000 unique members.

[2] The plots were developed using a Social Network Analysis tool called Pajek.

See: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

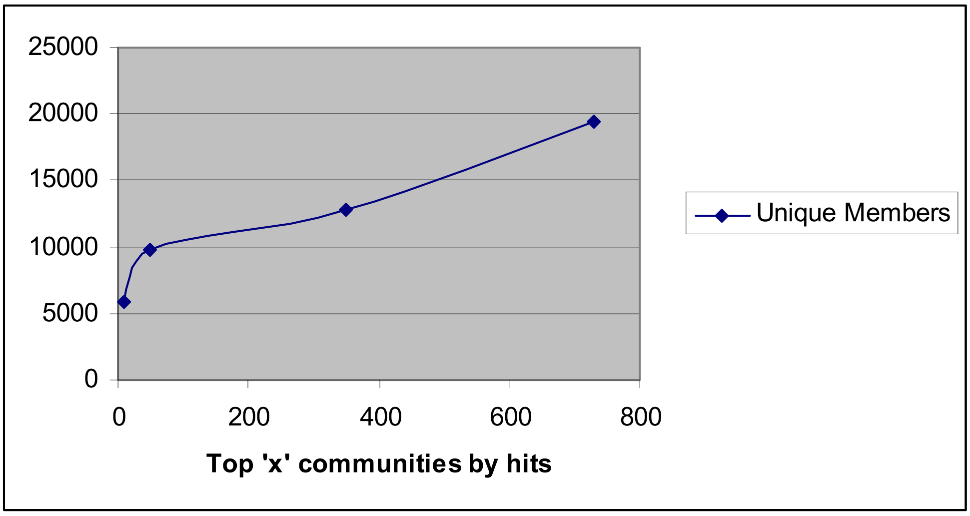

Figure 13 – Community Membership Statistics

2.1.1 What does CSC's interest profile look like?

The following map identifies CSC's core communities. The relationships are indicated by the degree of common membership between the communities. For example, one can see the relationship between the Windows 2000, IT security and Networks communities in the top right. The size of the nodes reflect the number of linkages to other communities. One could infer a CSC core competency around project and programme management, java and web development, catalyst, e-business, architecture and design and Windows 2000 from this map. The map below only includes communities that have a common membership of 200 or greater. The Project and Programme community and the Catalyst community share the most common members at 540.

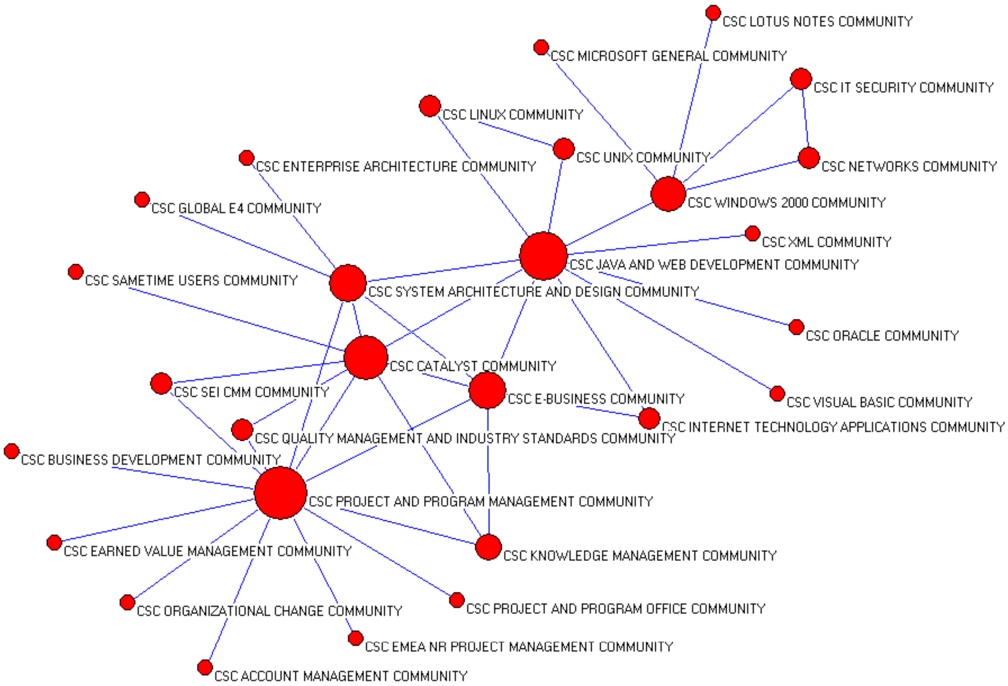

Figure 14 – CSC's Core Community Relationships

The other use of the information is to alert potential members of other communities that they might also be interested in joining. To take a leaf out of Amazon.com, prospective members could be informed with something like "thank you for joining the Networks Community, members who have joined this community have often also joined the Windows 2000 and IT Security communities".

2.1.2 Who has the potential to be a key knowledge broker?

By assuming that the communities that a staff member joins is a reflection of their own interest profile, we can analyse the network to find those staff members who share a common interest profile. This is done by determining how many community memberships that staff members share in common. For example if Joe and Sally have 6 common community memberships then a link would be established between Joe and Sally with a link value of 6. The following maps show the relationship network that exist for those members with the highest shared membership link values. Clearly the more communities one joins, the more chance one has of being on this map. But it would also indicate a broad range of interests and therefore increased potential for being a good knowledge broker. Of course the maps can say nothing about the individual's personal qualities with respect to knowledge broking, they can only tell us that these people, based on their interest profiles, are structurally well placed to act as knowledge brokers across the organisation.

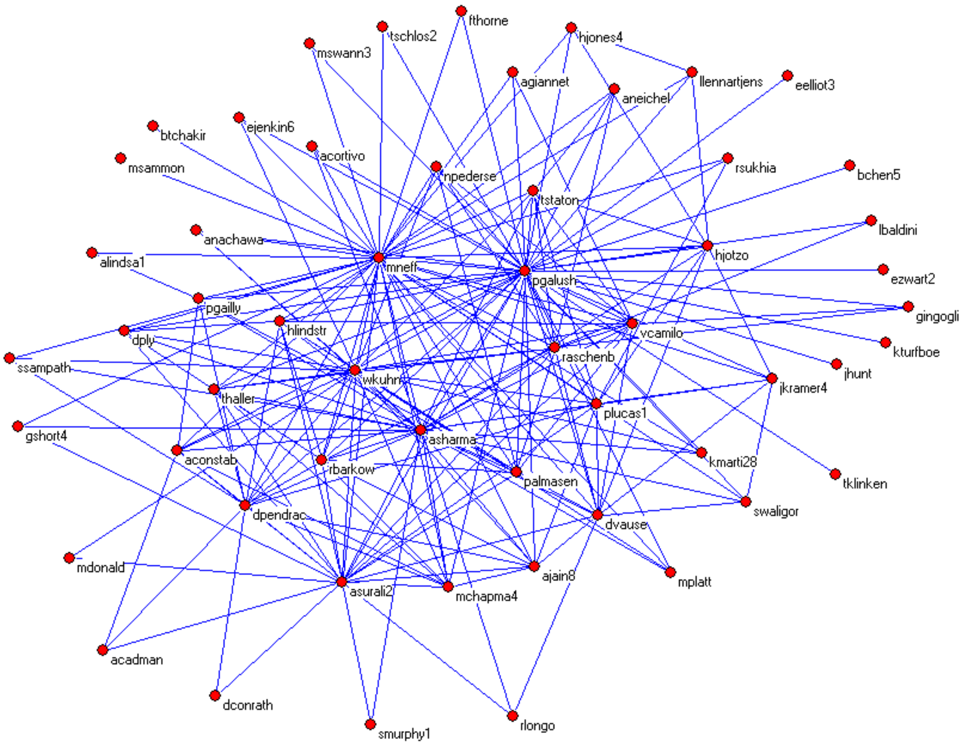

Figure 15 – Potential Key Knowledge Broker Network

2.1.3 Regional Patterns within the Broker Network

Can we see any geographic patterns in the broker network? Member data was used to classify members according to 8 regions. Some 13,000 members of the top 350 communities were included in this sample. The USA was roughly split into North East, South and West. The other regions were UK, Europe, Scandinavia, Canada and Asia Pacific. The following map shows the broker network for the top 116 brokers.

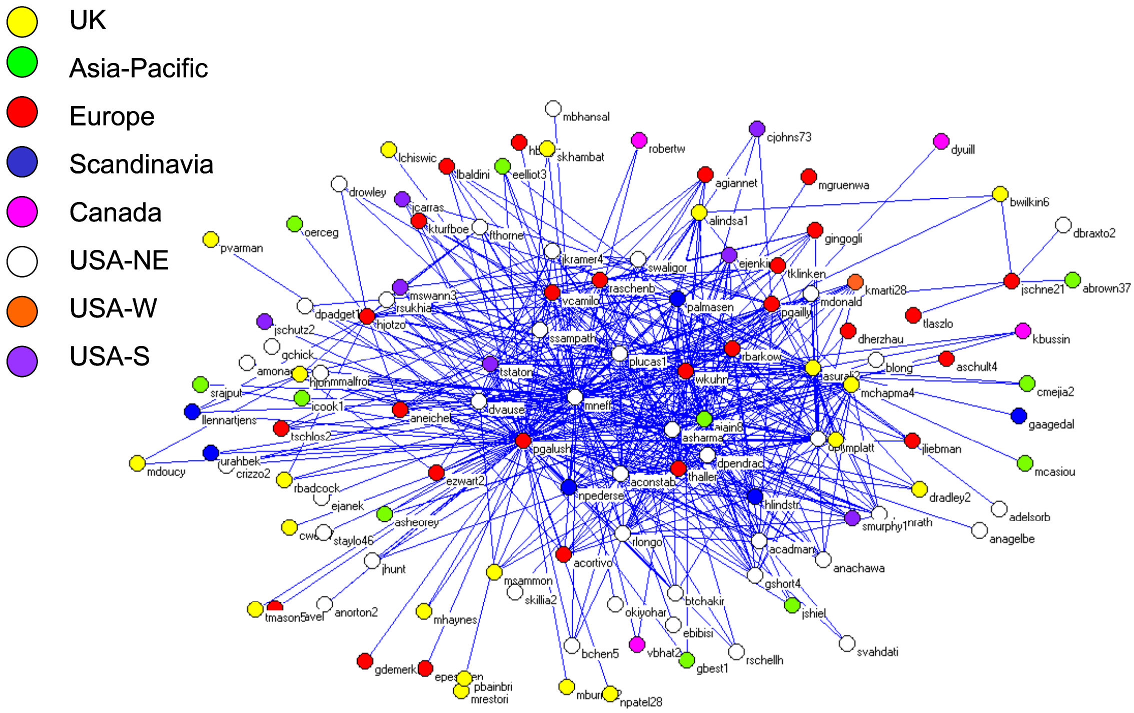

Figure 16 – Top 100 Broker Network by Region

USA North East, Europe and Scandinavia appear to dominate the core of the network. Why might this be so? The UK, Asia-Pacific and other regions appear to be more peripheral. No significant regional clusters are evident, indicating a fairly even spread of interests. Other possible interpretations are left to the reader.

2.2 Selected Published Case Studies

This section provides a selection of web mining case studies that illustrate some typical examples of the application of web mining technology. A unique feature of the selected case studies is that I didn't do the summarising! The Text Analyst product from megaputer (www.megaputer.com) was used to summarise the text from the original documents. The summaries are between 15 and 30% of the original document size. Only modest editorial adjustments were made, mostly with formatting related to document conversions from PDFs prior to feeding the texts into the summariser.

As you will see, the technology does a reasonable job. During my experimentations I found that the technology worked best with reasonably large documents that provided a reasonable textual flow, rather than short dot point style documents.

2.2.1 Predictive Modelling in Insurance[3]

Modeling in Insurance By Marty Ellingsworth and Dan Sullivan

Fireman's Fund Insurance Company tried a wide range of analytic techniques to understand rising homeowner claims and suspicious auto claims, but could not find predictive patterns in the data. The insurance company's team of analysts, led by one of the authors (Ellingsworth), realized the problem was not with their techniques, but with their data. To satisfy the accuracy needs of the modeling programs, the company used basic text mining techniques to isolate new attributes from the text and then combined those with previously available structured data to expand the total amount of relevant usable information.

Fireman's Fund subsequently discovered that success might just mean paying closer attention to the supply chain of information where basic data features originate.   In this article, we will describe a basic text mining technique, term extraction, and discuss how it was successfully used at Fireman's Fund to gain insights into urgent business problems.

Like all text mining techniques, this one maps information from unstructured data into a structured format. By mapping the free-form text to a feature vector, the text is modeled in the same attribute/value model used by structured data and thus lends itself to analysis using traditional business intelligence tools such as ad hoc reports, OLAP analysis, data mining and predictive modeling. When suits were brought against Firestone for faulty SUV tires, Fireman's Fund turned to free-form text analysis to determine if any of their claims related to the

[3] DM Review, July 2003

litigation. Fireman's Fund Matches Techniques to Problems Mastering information is a critical competency for success in the insurance industry.

Marty Ellingsworth is the director of operations research, a division of the Customer Research and Strategies group, in marketing at Fireman's Fund Insurance Company.

2.2.2 One-Line Market Research[4]

BLUE MARTINI SOFTWARE CUSTOMER CASE STUDY

Blue Martini Business Intelligence At Work:

Charting the Wilderness of MEC Website Data Blue Martini Software Customer Case Study - 1
Key Insights:

• The further away customers live from an MEC retail store, the more money they spent
• Analysis of customer postal codes generated a geographical heat map that revealed "hot" areas for future physical store sites
•A change from free shipping to flat-fee shipping resulted in a 6.5 percent increase in total sales and 18 percent increase in average sales per order
• Customers who increased their purchase level from under $200 to over $200 accounted for 8.4 percent of the revenues and had unique identifying characteristics, which will help MEC target future migrators
• Robot/crawler removal revealed that the average human session was 26 percent longer than the average session. “We felt that the analysis was very useful and intend to act on what we learned very quickly.

A GEOGRAPHICAL BREAKDOWN OF MEC CUSTOMERS

By overlaying Canadian and US latitude-longitude data with customer postal codes, Blue Martini presented a geographical breakdown of MEC customers in Canada and the US. Most customers were clustered in regions around the seven MEC retail stores.

MEC provided BMAS with information from its Web site databases, including clickstreams, transactions, content, product catalog, and customer information. The Blue Martini team analyzed the MEC data using Blue Martini’s Business Intelligence suite and presented a detailed set of findings to MEC. The presentation featured key insights across a broad range of topics:

•From the geographical distribution of current customers to the characteristics of emerging customer segments
•From the performance of top-selling products to the predicted performance of yet-untried product combinations

[4] Case Study provided by Blue Martini Software

•From the impact of marketing campaigns to the effectiveness of online search
•From the online browsing behavior of MEC site visitors to the effectiveness of certain pages (or regions) of the site.

The following sections present some of the key insights gleaned from the MEC data using Blue Martini's Business Intelligence suite. Moreover, if the last search in a session was successful (i.e., results were returned), that visit generated almost three times as much revenue as visits that included no search at all.

• About one in thirty failed searches used one of six keywords
• Revenue per session that included a view of a "For Your Information" text, providing additional product information, was nearly six times higher than the average session
• Product affinity analysis identified significant cross-sell opportunities that were not currently offered on the site
•A portion of the site dedicated to visitors advertising used products for sale (Gear Swap) was an island.

The "bot" filtration algorithm included in the Blue Martini Business Intelligence tools ensures that only true human traffic will be analyzed.

Jest aside, market basket analysis has been used extensively by retailers to determine effective cross-sells. The association algorithm in Blue Martini's Business Intelligence suite identified over 160 significant product associations for MEC. Blue Martini was then able to make enhanced cross-sell recommendations based on the automatically generated association rules to complement the cross-sell recommendations already in place on the MEC website.

I N SEARCH OF SUCCESS : ANALYZING SEARCH EFFECTIVENESS

In addition to the wealth of data and analysis prepared for MEC, the Blue Martini team also showed the retailer how one basic feature, the search engine, can significantly affect site performance and revenue. From extensive research on Web sites of companies across different industries, Blue Martini has proven that successful searches convert more visitors into purchasing customers.

M OVING FORWARD WITH BLUE MARTINI BUSINESS INTELLIGENCE

Blue Martini made several recommendations to MEC that go beyond the standard operation of the MEC website. The key recommendations included:

• Consider a tiered shipping charge based on order amount (for example, free shipping on orders above $100 CAD)
• Implement tighter interconnection between MEC shopping and the Gear Swap section.
MEC is now implementing this recommendation
• Use the Blue Martini's Web site effectiveness measurement functionality to guide site redesign

• Complement existing product recommendations with those automatically identified using association rules (market basket analysis)
• Use the online site as an experimental laboratory to try out new product offerings and study customer response to policy changes or marketing promotions
• Continue to invest in the search engine and monitor failed searches for new trends or changes in customer preferences
• Combine online data with offline data from retail stores to obtain a 360 degree view of the MEC customer
• Consider adding highly searched products to the product offering The project presents MEC with a comprehensive report card of its online operations and empowers the MEC management team to make well-informed decisions on boosting online sales and enhancing customer satisfaction.

2.2.3 Identifying occupational health hazards from call centre notes[5]

PolyAnalyst provides the user with unique tools for analyzing data in textual format by extracting key concepts from text stored in a database format, categorizing individual database records, and deriving from text quantitative knowledge in a format that can be exploited by other machine learning engines of this data mining system.

This is the first Tutorial illustrating the unique text analysis capabilities of PolyAnalyst. Traditional statistical and data mining techniques provide little help in this case, and PolyAnalyst offers novel and unique text mining capabilities that help the user address the task.

<u>Using Text Analysis exploration engine</u>

The exploration engine works for a couple of minutes and returns a collection of term rules (displayed in the Project Tree in a folder named TA_process_name) that can be used later for categorizing individual records, and a textual report summarizing the findings of the system.

First, we obtain the result through the more complex but thorough comparison using the Summary Statistics exploration engine. Run Summary Statistics exploration engine, leaving only the Localization field in the analysis, on the World dataset, and then - on Cause_fall_1 and Cause_lift_1 datasets, representing those people who fall and lift something, correspondingly.

Comparing the result of these explorations, one might see that groups of people who fall and lift something have significantly different distributions of trauma localization than the overall data. The overall distribution of localizations of trauma:
Localization Frequency Percentage
Lower Back Area (Lumbar and lumbo-Sacral Muscles) 163 18.76%
Finger(s) 87 10.01%

[5] This case was presented as a PolyAnalyst tutorial by Megaputer

Knee 64 7.37%
Hand 50 5.75%
Wrist 39 4.49%
Upper Arm 38 4.37%
Ankle 32 3.68% .

The distribution of localizations of trauma for people who fall: Localization Frequency Percentage
Lower Back Area (Lumbar and lumbo-Sacral Muscles) 17 12.32%
Knee 14 10.14%
Wrist 10 7.25%
Foot 10 7.25%
Multiple Body Parts and Body Systems (External and Internal) 8 5.80%
Lower Leg 8 5.80%
Multiple Head Injury 7 5.07%
Upper Arm 7 5.07%
Finger(s) 6 4.35%
Ankle 6 4.35%.

The distribution of localizations of trauma for people who lift: Localization Frequency Percentage
Lower Back Area (Lumbar and lumbo-Sacral Muscles) 35 33.98%
Upper Arm 7 6.80%
Shoulder(s)(Armpit, rotator Cuff, Clavicle, Scapula, Trapezius) 7 6.80%
Pelvis 6 5.83%
Abdomen (Including Groin) 5 4.85%
Finger(s) 5 4.85%
Disc (Trunk) 4 3.88%
Multiple Trunk 4 3.88%
Multiple Body Parts and Body Systems (External and Internal) 4 3.88%.

Thus far, we had to compare the correlations between various causes of injures and the corresponding localizations of trauma indirectly by running the Summary Statistics exploration engine separately on each resulting dataset and then comparing the results with each other. PolyAnalyst provides an additional data analysis technique, Link Chart, which allows the user to obtain and visualize relevant results in just a few mouseclicks.

2.3 Internet/Market Case Study

The following examples have been developed to demonstrate what's possible with web mining when applied to business/marketing intelligence applications, using the web as the primary information source. The Vivisimo (www.vivisimo.com) meta-search and classification engine has been used as the web mining tool for these demonstrators. Vivisimo uses heuristic hierarchical clustering algorithms to group textual documents into a hierarchy or tree, providing a textual label for each node in the tree. The closest parallel to traditional data mining techniques is the rule induction methods described within the machine learning literature (Quinlan, 1992). These decision tree / rule induction techniques mine "human understandable" decision trees from structured repositories of both text and numeric data. For examples described below, there was no capacity to manipulate or tune the clustering algorithms used by Vivisimo, though one could impact what unstructured sources were searched e.g. could be the whole of the web or just a particular web site.

2.3.1 Exploring the BPO Market Example

This first example looks to characterise a market that CSC has a strong interest in i.e. the business process outsourcing market. The methodology used was to search the web using Vivisimo using a search string like "CSC and BPO or Business Process Outsourcing". Vivisimo then produces a hierarchy of elements categorising the hits it achieves form the search engines that it employs. On the left hand side of the screen the hierarchy can be seen and expanded as required. This search was completed for a number of competitors in the BPO market, producing a hierarchy for each search. From this collection of textual hierarchies a set of common labels or concepts were selected that provided the highest number of hits. It should be noted that Vivisimo provides an option to cap the number of hit to a selected limit. To ensure that all competitors were evaluated equally, only the top 200 or so hits for each competitor were used. These concepts were then plotted on a bar graph showing which concepts were most prominent (based on hits/concept) for each competitor, in effect showing how their top 200 hits were distributed amongst the selected concepts.

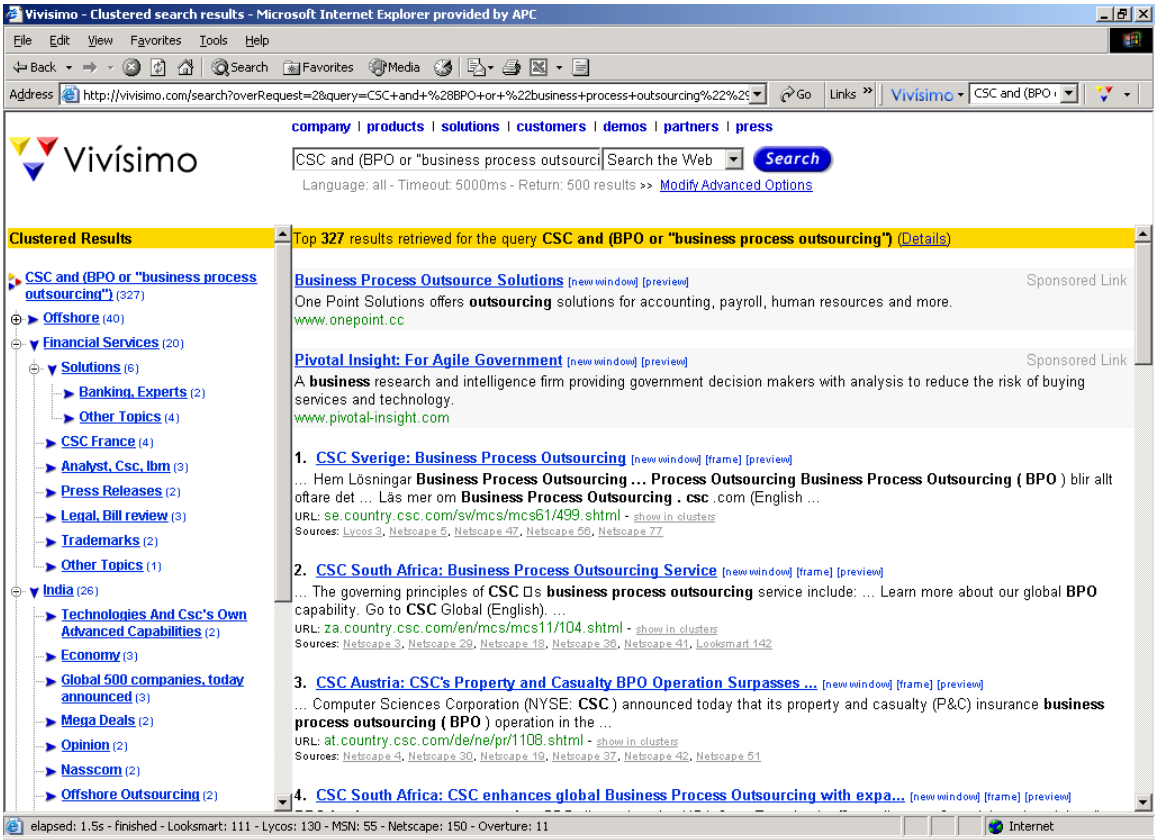

Figure 17 – Vivisimo BPO Search

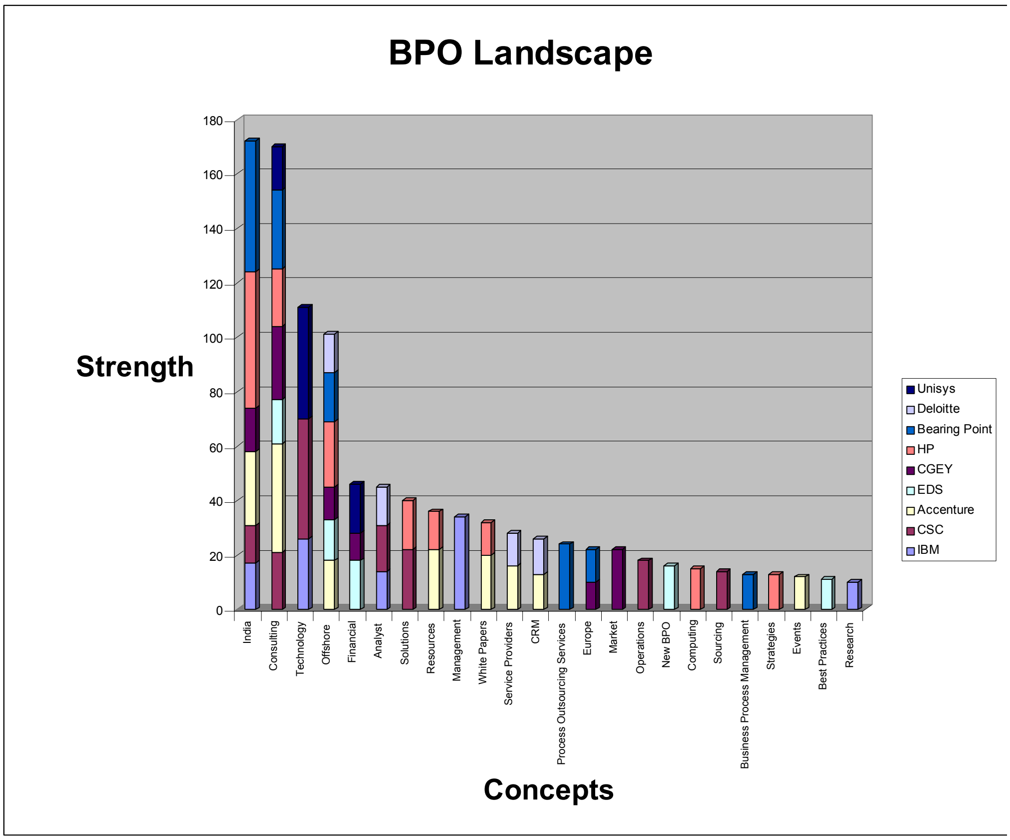

Figure 18 – BPO Market

The BPO market landscape tells us that India and Consulting are the strongest concepts in the BPO market followed by Technology and Offshore (which is clearly related to India). Of the competitors one can see that Bearing Point and HP look to be the most aggressive in the India/Offshore strategy for BPO. CSC is also a player in the India and Consulting areas but appears to be strongest when it comes to technology. Given that the source for this analysis is the whole web, one could infer that this graph is a reflection of the public perception of how each player is operating within the BPO market.

Another feature of the plot is the number of concepts that aren't shared amongst other BPO competitors. 50% of the selected concepts had only a single competitor associated with them. For example, Operations and Sourcing seem to be unique BPO concepts for CSC. This number may also reflect the immaturity of the marketplace for BPO.

2.3.2 Other Market Examples: CRM and Mobile Phones

Continuing on with the same methodology, two other markets were explored viz. Customer Relationship Marketing (CRM) and the Mobile Telephone Market. These two markets were selected because they both have a dominant competitor in Siebel (CRM) and Nokia (Mobile Phones) and markets that are more mature than BPO. The following plots provide the results of these analyses:

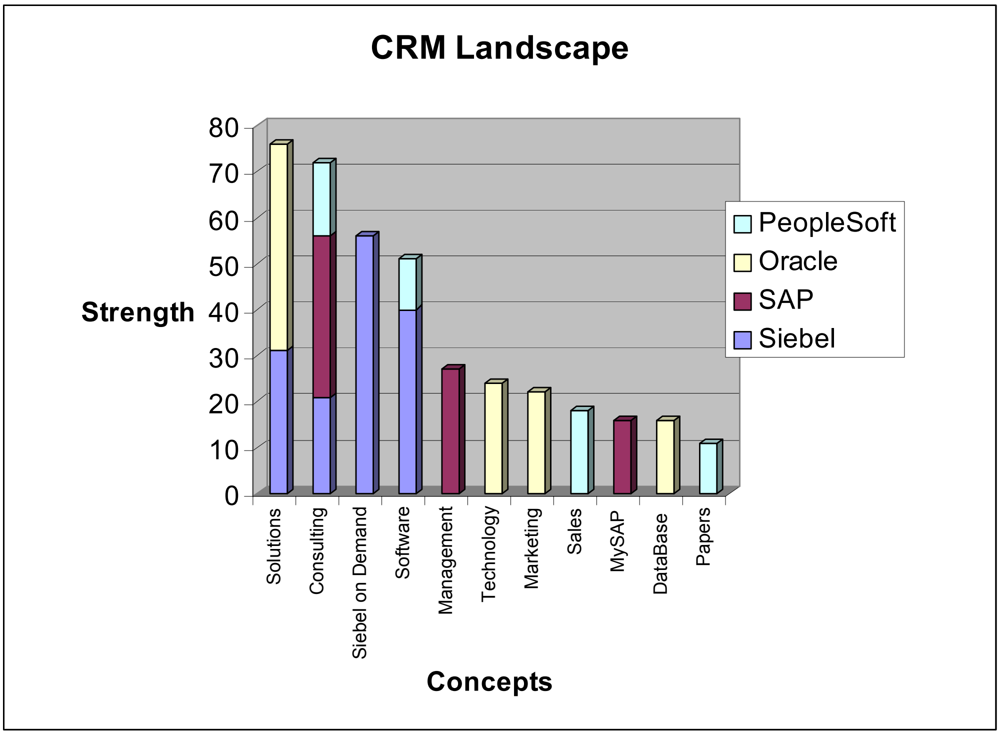

Figure 19 – CRM Market

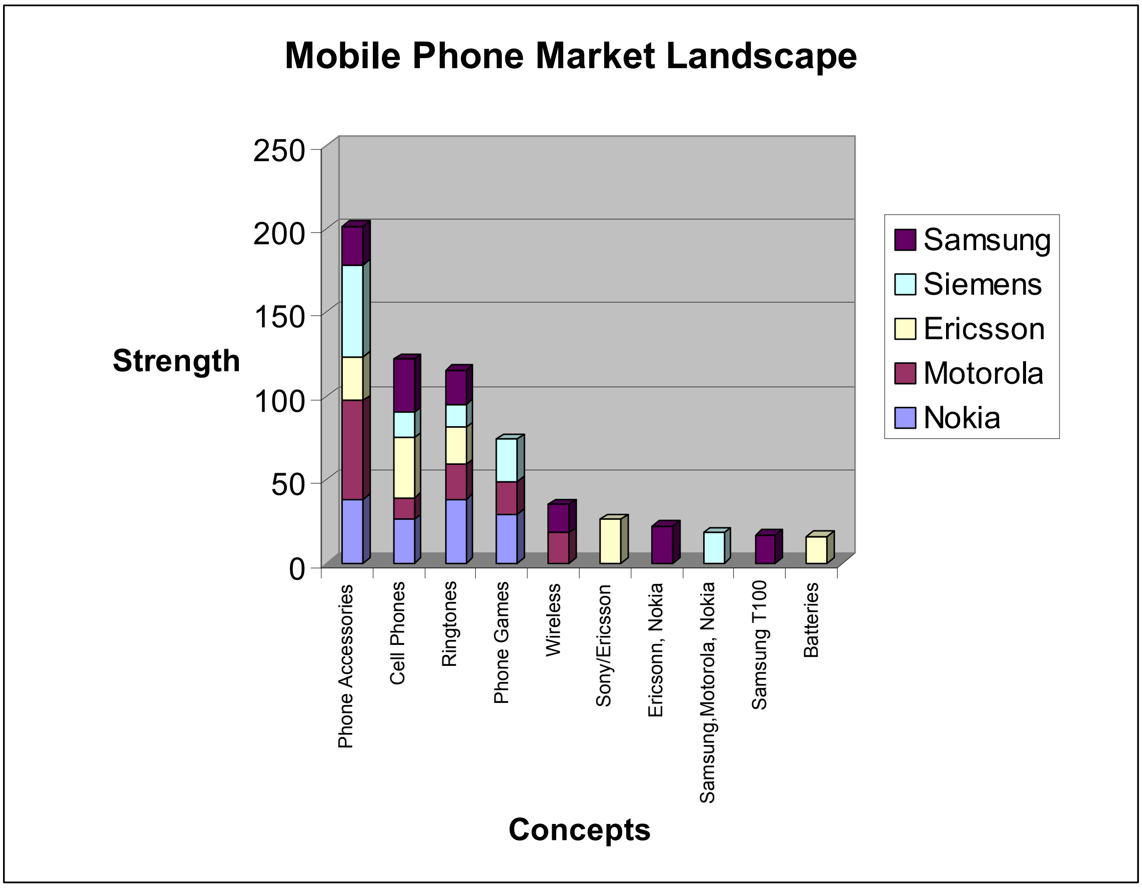

Figure 20 – Mobile Phone Market

What is common from the above plots is that the dominant market leader is concentrating only on the four leading concepts for each market. In the CRM market, Siebel is very focussed on Solutions, Consulting and Software as well as strongly promoting its Siebel on Demand concept. Siebel's competitors like Oracle and SAP are also promoting unique messages but at this stage perhaps not as successfully.

The mobile phone market appears to be driven by accessories, followed by the promotion of cell phones themselves and then ringtones and phone games. Nokia is focussed on the top four concepts while the market number two, Motorola has a similar profile.

So what is the value of this sort of analysis? Clearly these analyses are only of value if they are consistent with the beliefs of expert market commentators. This test hasn't been done but the results are left to the reader to judge. If in fact they are a good reflection then web mining can add significant value to market analyses. The above analyses can be conducted in just minutes, rather than weeks or months.

2.3.3 The IT Services Alliances Landscape

This example provides an insight into "who's working with who" in the IT services market. Alliances and partnerships are commonplace in the IT services market. Knowing the nature of alliances across the industry can be a key input into any company's competitive strategies. In exploring the web for alliance information, there appeared no single source for capturing alliance information for the IT services industry. Going to individual company web sites could provide a suite of alliance announcements, but rarely in any managed fashion. Also, there existed information about alliances that were not evident on company web sites. Searching for all evidence of alliance activity across the web proved quite tedious, with many announcements replicated in many different areas and little consistency between what could be found by a search using a company's internet web site and a generic search engine searching the site.

To reduce the amount of information to a manageable level, the Vivisimo tool was once again used to search the web for alliance information for a number of competitors. Again the search was limited to the top 500 hits. The hierarchy developed was then analysed to identify where company names were identified and at what level of the hierarchy. A scoring scheme was established to weight the alliance strength based on the number of hits that the alliance partner was identified with. The score was down rated slightly if it only occurred at the lower levels of the hierarchy. A measure for "alliance intensity" for each competitor was devised by totalling the number of hits that related to a specific alliance company from their top 500 hits, that mentioned alliances. This could be cautiously interpreted as the amount each competitor is reliant on alliance partnerships.

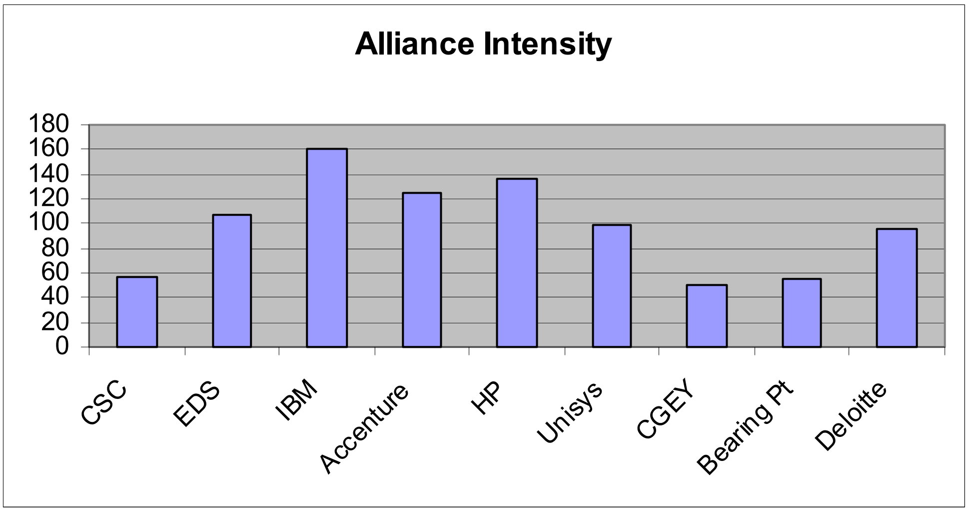

Figure 21 – IT Services Alliance Intensity

This plot identifies IBM and HP as being the most alliance intensive. This may reflect the fact that both companies also provide hardware and systems software, which could not easily be separated out for this analysis. The CSC rating suffers to some extent through CSC's practice of branding large outrsourcing alliances like Prime, Eagle, Pinnacle etc., which largely hides the individual participants in these alliances.

To see how each of the major competitors related to each other and other suppliers in the market place, the data was then fed into a network mapping software (Pajek) to help gain a picture of the dynamics of the alliance activity.

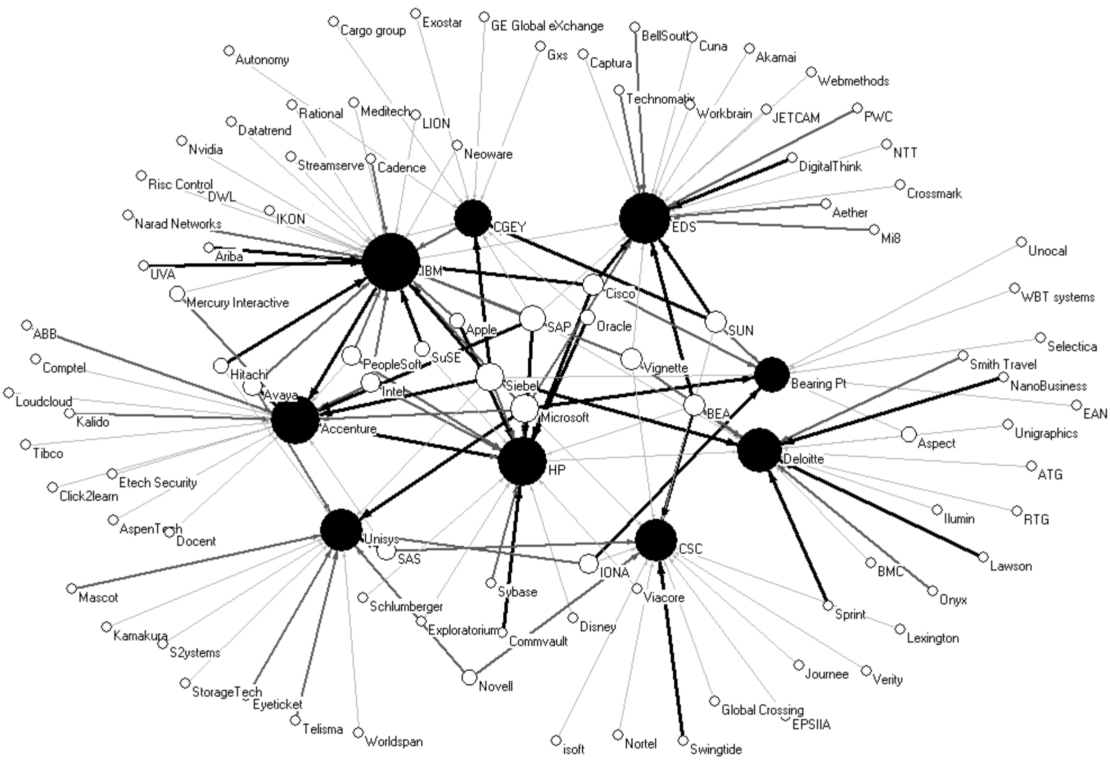

Figure 22 – Alliance Network Map

The alliance map show each of the competitors surveyed as black circles and their alliance partners as open circles. The size of the circles reflect the number of connections that each node has. For example, IBM has the largest number of connections and therefore is the largest node. The surveyed companies are naturally larger than non surveyed companies (open circles), who's size reflects the number of alliances they might have with the nine surveyed companies. The thickness of the line reflects the strength weighting (based on number of hits and position in the hierarchy). Note that the map does not show all of the alliances that a company may have. An individual search for CSC identified over 100 alliances. Vivisimo limits the search to the top 500 hits of pages mentioning CSC and alliances. While this is not ideal, the nature of the market from what was analysed identifies less than 20 "core suppliers" to the IT services industry. The rest of the alliances appear to be one-off unique alliances that the individual companies may have formed to perhaps joint venture on a contract or bid requiring more than purely IT services, pursue a niche strategy or exploring a new business opportunity.

The spatial layout of the map above is also important. Essentially the network layout algorithm attempts to minimise link distances. The relative distance of suppliers and alliance partners from the nine surveyed companies can loosely give some indication of the relative closeness of the relationships. You might want to liken the network map to a city street directory map. The core suppliers all appear in the "central business district", the individual suppliers who might link to only a single IT service provider are in the "outer suburbs". The map appears to show CSC is further removed from the core suppliers than its key competitors. Does this match with the general perception? Again the quality of the information can only be judged by how consistent it is with expert opinion.

Another way of trying to authenticate the above alliance map is to go to the web sites for some of the vendors and look to see how often the surveyed companies are mentioned. For example, if we are trying to authenticate Siebel's strong links with Deloitte, Accenture and IBM versus CSC's weaker link we would find 39, 58 and 135 mentions respectively verses CSC's 12 mentions. Similarly, Microsoft is seen to have strong links to Bearing Point, Unisys and Accenture while again CSC has only a weaker link. From the Microsoft site of partner case studies CSC has 5 cases mentioned versus 18, 20 and 20 for Bearing Point, Unisys and Accenture respectively.

The qualitative nature of these measures would indicate that the level of authentication can only be improved by data triangulation i.e. looking for evidence from different data sources to confirm or otherwise the hypothesis, as illustrated above. While the text analysis algorithms improve on simple word or document counts, they will still fall short of full text understanding. A case in point is the lack of an alliance link between Accenture and BEA. While there was a lower number of web references to such an alliance, this particular statement occurred in one of the published documents : *"Accenture is our biggest integration partner," says Alfred Chuang, CEO of BEA"*, emphasising the fact that we still have some way to go to achieving full text understanding from textual analysis techniques.

Despite the above shortcoming, the alliance map does appear to provide a good "first cut" of the overall alliance landscape of the IT services industry. Where the data does not fit well with perception, further focussed research can be conducted to either authenticate or refute the linkages shown.

3 DISCUSSION ON IMPACTS AND OPPORTUNITIES

This section reports the results of a number of focus group meetings held to help articulate what the opportunities might be for Web Mining.

3.1 Current Situation / Market

The market for Web Mining can be characterised by both the attributes of a good application prospect and the early adopter industries.

The important attributes of a good Web Mining prospect are:

- The existence of large electronic textual repositories

Organisations that typically store large amounts of text are good candidates for extracting more value from the resource. Typically, news services, government policy stores, legal case files, patent stores, police or military records are examples of these.

- Humans currently undertaking tedious processing of text

Areas where people are being asked to review or process large amount of text, are prone to fatigue and failure. Tasks like reviewing applications or statements, literature searching, developing alerts from news feeds are all text processing intensive.

- Electronic usage logs available

As commerce moves increasingly on-line, the opportunity to monitor and mine usage patterns increases. Usage logs are still only being lightly mined. With the advent of web services and service oriented architectures, the potential for usage log analyses can only increase.

- Patterns are important

Where important patterns of either usage or in content are effectively hidden from conventional access methods i.e. where the data base query doesn't tell you enough. Traditional data mining applications like fraud detection, criminal networks, subtle buying patterns, which see multiple factors interrelate, are examples.

The Early Adopter Industries appear to be:

- e-Retail

Basically those industries doing commerce on the web have been the early adopters of web mining for usage analysis.

- Insurance

The Insurance industry provides multiple opportunities. The complex web of sales agents and insurance providers provides an opportunity to discover patterns of policy / agent / client relationships. Several insurance companies are now providing web sites for their agents, leading to usage pattern analysis. Searching for new product opportunities is another potential application. Products are quick to develop but do they overlap too much with other offerings or others in the marketplace?

- e-Government (Border Protection)

A number of applications have been identified in relation to border protection e.g. identifying drug traffickers or terrorist networks from synthesising news articles and / or intelligence reports.

- Defence

Similar applications to border protection, but with a military context.

- Life Sciences

Driven by the need to explore and review large amounts of textual reports to identify new opportunities for genetic and biological developments. In a related development CSC has assisted the US Patent and Trademark Office in assessing genetic patent applications. CSC used artificial intelligence techniques to interpret the "genomic language" in the same way as computers have been used to understand natural language (Saunders, R, 2004).

The source data for web mining is increasing at a far greater rate that one can process it. Even Google is only indexing a fraction of the unstructured data available on the Internet. One could view the world of unstructured information as looking something like:

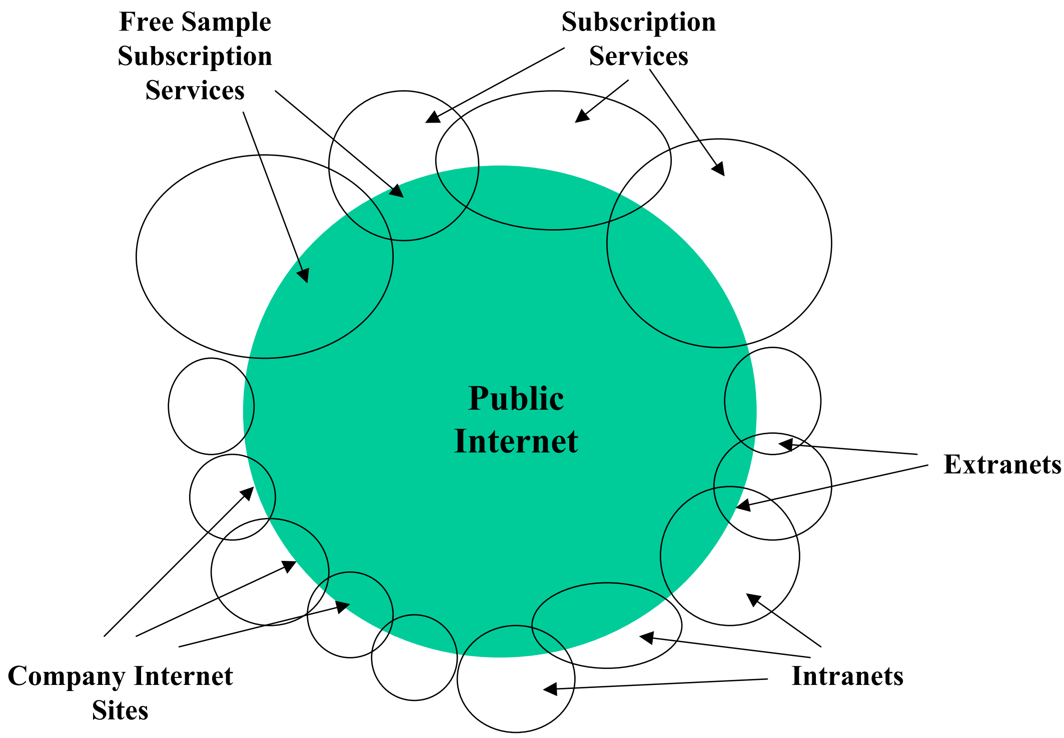

Figure 23 – The World of Unstructured Information

Clearly, the major source of information is the publicly accessible Internet. Additional to the Internet we have subscription services like Reuters, Bloomberg, Forbes and specialist services like D&B, all with a partial Internet presence but basically secured to paying clients. Company Internet sites usually provide a subset of information provided on their Intranet. Extranets secured between companies is another web based resource. Ultimately, Web Mining applications will be limited to the degree of access that can be granted to secured repositories. For company intranets and extranets access would need to be negotiated with each company. Companies themselves may choose to employ web mining of their intranet for their own internal purposes. Web mining provides an avenue for web masters to derive more value from their usage logs by more accurately profiling their client base. For subscriber sites, the opportunity may need to be one of partnership i.e. how web mining could enable the service provider to draw more value form the information asset they have and presumably on-sell this to their client base.

3.2 Potential Value Adding Services Identified

A synthesis of the focus group discussions centred on two major applications themes, Smart Alarms and Product / Market Intelligence:

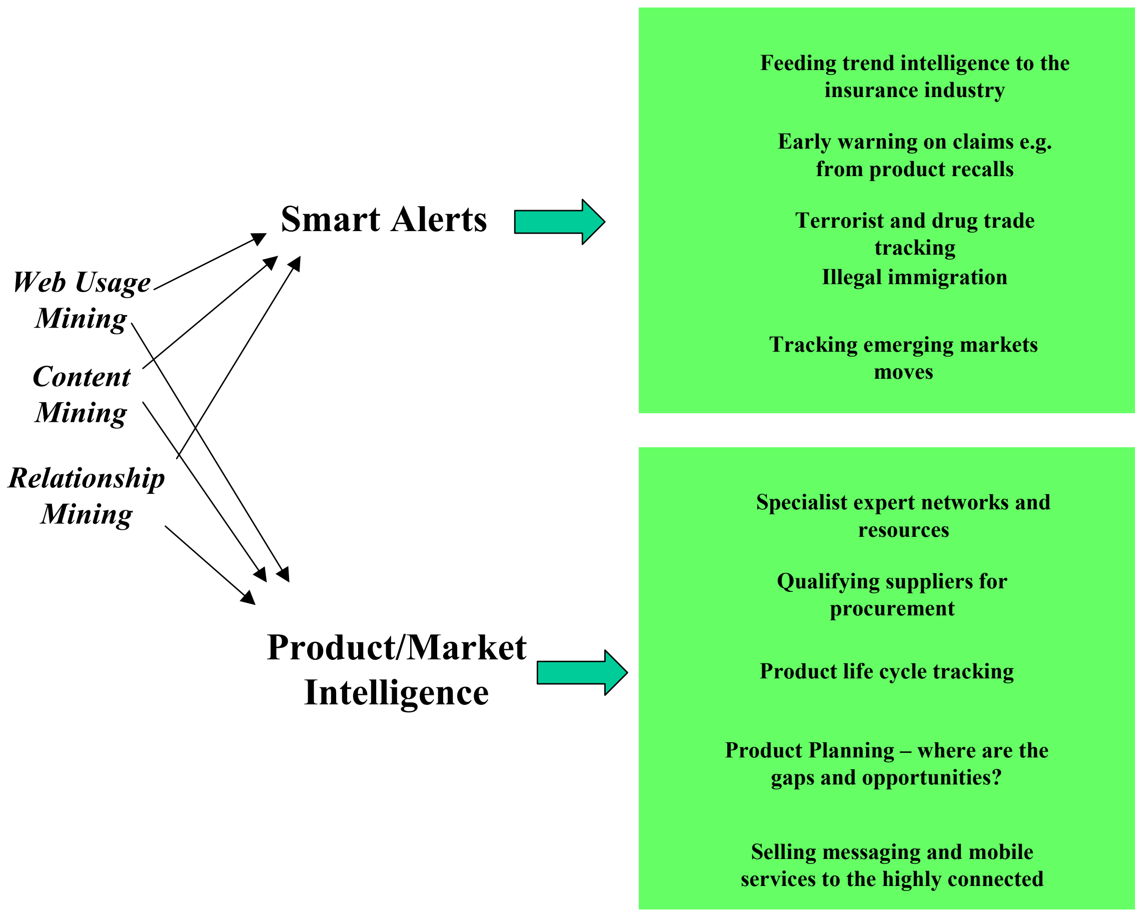

Figure 24 – New Service Themes

These new service themes are derived from either content mining, and web usage log mining and the emerging ability to discover relationship patterns from web based data. The themes are Smart Alarms and Product / Market Intelligence.

3.2.1 Smart Alerts

As we can see from Figure 24 there are several opportunity themes collected under "Smart Alarms", ranging from terrorist or drug trade activity through to market place warning on insurance claims. Typically business intelligence staff would scan both public and private media for snippets of information that relate to areas of interest for the organisation. These snippets might be distributed broadly to people who may have expressed an interest or subscribed to a particular alert service. What the individuals do with these alerts will be contextual, but it is likely that those receiving the alerts will use them to help frame a mental picture as to "what's going on". Are there any trends developing? does this new information strengthen or weaken some prior assumptions?, does the alert generate the need for a new line of enquiry? The mental model will change over time as new alerts come in.

Smart alerts are about taking the next step in business intelligence alerts by helping the end user build and interpret the evolving situation. This will mean integrating new intelligence with past intelligence to represent how patterns or trends are evolving over time. By using visualisation technologies to see the patterns as they emerge, the smart alert is able to highlight key trends mined from the data and alert the intelligence consumer in the context of what has been happening in the past. In essence the smart alert systems would be continuously mining web based repositories and using visualisation techniques to illustrate emerging trends or developments.

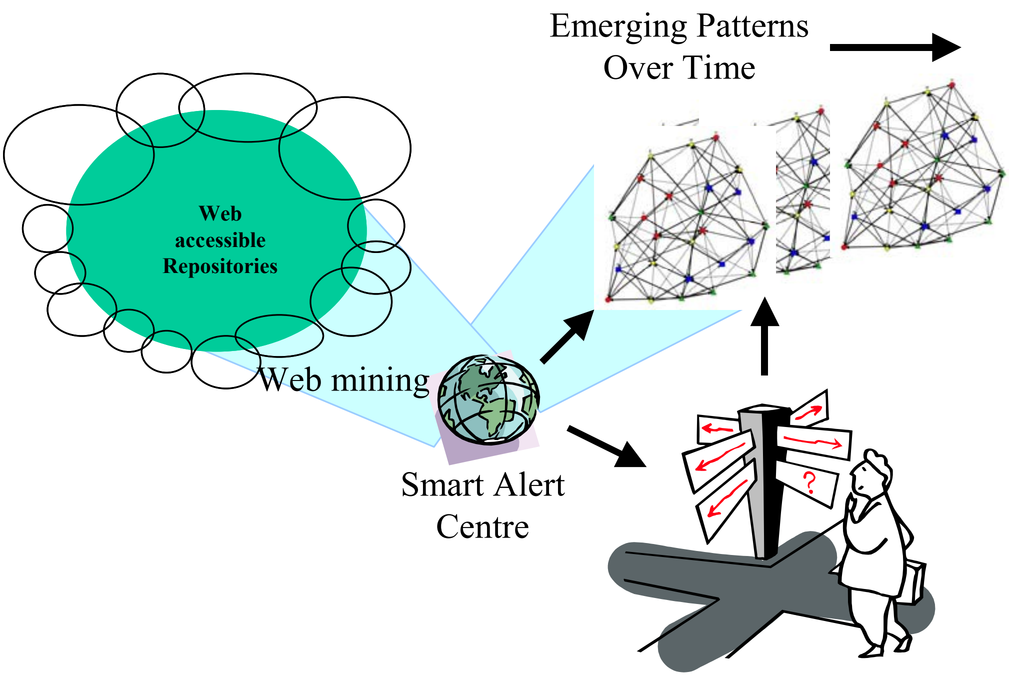

Figure 25 – Smart Alert

The intelligence consumers are presented not just with a textual alert but a representation of trends and patterns mined from the web over time. For example, a customs department may be vitally interested in drug trading patterns that are evolving over time. Today, intelligence staff will clip articles and stories that they perceive are of interest to their subscribers and send them off. These stories may contain information about the type of drug, the source countries, known traffickers etc.. Tomorrow, with the smart alert, the news articles will have been mined to extract identified key concepts like source countries, known traffickers and drug types. The frequency of occurrence and the diversity of sources from which the intelligence is drawn, will be used to update the patterns and trends previously collected. The news alerts would be distributed along with updated trend patterns or emerging relationships developments, providing the intelligence consumer with a "high fidelity picture" of the emerging situation. Additionally, it will bring all subscribers to the intelligence up to the same level of understanding, in the domains of interest.

3.2.2 Product / Market Intelligence

The potential applications in this area ranges from providing assistance at the industry sector level on expertise networks, finding new product opportunities or new sales opportunities from information that can be mined from the web.

An example of a potential expert network is with the oil and gas industry. In this industry, exploration and production projects are often developed as joint ventures. Over time the industry becomes highly networked as alliances and partnerships develop. Research and technical developments are often published in the public domain, sometimes facilitated by professional bodies. Web mining can be used to identify the nature of the expert networks that exist by analysing information from on-line forums, technical paper repositories and company directories.

In industries where new product offerings are non physical e.g. an insurance or investment product, the ability to spot a gap in the market place to exploit is becoming harder. Web mining can be used to explore products available on the market and described on company web sites. As much of the product description is textual, web mining can be used to extract the key concepts from the product offerings to build a picture of the current market (similar to Figure 20 – Mobile Phone Market). Product managers can quickly see what pitch their competitors are going to market with and be able to devise a strategy to either compete or complement.

Another market related opportunity might be to identify potential clients for a particular service, based on their behaviour patterns as captured on the web. For example, if a market level analysis shows a particular client is heavily connected to multiple suppliers, we might assume that they would be in need of a procurement system that can cater for a large number of suppliers. This same information might also be used to qualify a potential supplier i.e. what other companies are working with this supplier? Does the evidence fit the supplier claims?

4 SUMMARY AND CONCLUSIONS

This report has investigated the important opportunity of how we can gain business advantage from the world's greatest information resource, the World Wide Web. It has been established that between 80% and 90% of the information that we use in our day-to-day work exists in unstructured, mainly textual repositories. Many of these repositories are now accessible through web technologies, be it an internal intranet or the Internet. This situation will only grow to the point that nearly all the information that we work with will be web accessible.

Web mining is a technology, application and product suite that is aimed at discovering valuable business insights from the vast information repositories accessible through the web. Successful applications have been demonstrated in the areas of marketing and business intelligence, customer relationship management, biotechnical design applications, product design and positioning and knowledge management. A number of examples of web mining were provided illustrating applications like understanding the business process outsourcing market, analysing the network of business alliances in the IT services market, identifying where CSC's core interests lie and identifying potential key knowledge brokers for CSC world-wide.

Finally, a number of opportunity themes were identified. The development of "Smart Alerts" uses web mining to monitor patterns of activity over time. Used in an intelligence context, a smart alert will provide a high fidelity picture of emerging situations by visualising concepts as they develop from textual repositories.

Product and market intelligence opportunities can be enhanced through the use of web mining by providing a mechanism for assessing non-financial attributes of a product opportunity. Text mining technology can be used to characterise a marketplace by discovering key concepts adopted by competing firms. New products can therefore be positioned for maximum advantage.

Adar, E. and J. Tyler (2000). Zoomgraph, Information Dynamics Lab, HP Laboratories., Palo Alto, California, USA.

Bontis, N., M. Fearon, et al. (2003). "The e-flow audit: an evaluation of knowledge flow within and outside a high-tech firm." Journal of Knowledge Management **7**(1): pp.6-19.

Biggs, M. (2003). "Data Mining outside the firewall." InfoWorld Sept 5th 2003.

Butler, M. (2003). Autonomy: Automated Management of Unstructured Content, Butler group.(www.autonomy.com).

Chen, H. (2002). Trailblazing a Path Toward Knowledge and Transformation: E-Library, E-Government and E-Commerce, Dept. of Management Information Systems, University of Arizona, USA.

Chen, H. (2003). "Special Issue: "Web retrieval and mining"." Decision Support Systems **35**: pp.1-5.

Chen, Z., L. Wenyin, et al. (2001). "Web Mining for Web Image Retrieval." Journal of the American Society for Information Science and Technology **52**(10).

Cooley, R., B. Mobasher, et al. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97, Department of Computer Science, University of Minnesota.

Chuang, S.-L. and Chien, L.-F. (2003). "Automatic query taxonomy generation for information retrieval applications." Online Information Review **27**(4): pp.243-255.

Doherty, P. (2000). "Web Mining - the E-Tailer's Holy Grail." DM Review., January 28th.

Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "From Data Mining to Knowledge Discovery in Data Bases." AI Magazine Fall 1996.

Fung, G. P. C., J. X. Yu, et al. (2003). "Stock Prediction: Integrating Text Mining Approach using Real-Time News." IEEE CIFEr'03 Hong Kong: pp.395-401.

Gerdes, J. (2003). "EDGAR-Analyser: automating the analysis of corporate data contained in the SEC's EDGAR database." Decision Support Systems **35**: pp.7-29.

Getoor, L. (2002). Link Mining: A New Data Mining Challenge, Dept. of Computer Science, University of Maryland.

Gibson, D., J. Kleinberg, et al. (1998). Inferring Web Communities from Link Topology. ACM Conference on Hypertext and Hypermedia.

Hanish, D., J. Fluck, et al. (2003). Playing Biology's Name Game: Identifying Protein Names in Scientific Text. Pacific Symposium on Biocomputing, Kauai Marriott Resort and Beach Club.

Hearst, M. (1999). Untangling Text Data Mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Compuational Linguistics, University of Maryland.

Iyer, G., A. Miyazaki, et al. (2002). "Linking Web-based segmentation to pricing tactics." Journal of Product & brand Management **11**(5): pp.288-302.

Kanzler, J. (2002). "Buisness Intelligence Integration." Computer Technology Review **22**(12): pp.22.

Kohavi, R., L. Mason, et al. (2004). "Lessons and Challenges from Mining Retail E-Commerce Data." Machine Learning Journal, Special Issue on Data Mining Lessons.

Kou, Z. (2003). Extracting information from text and images for location proteomics. BIOKDD03: 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics.

Lock Lee, L. (2003). Does your community leave a digital footprint? Measures and metrics for communities of practice. KM Challenge 2003, Melbourne, Australia.

Manchester, P. (2001). "Every detail of the way potential customers use a website is recordable - and this information can be used to turn visitors into buyers." Financial Times, October 17th.

Menczer, F. (2003). "Complementing search engines with on-line Web mining agents." Decision Support Systems **35**(2).

Mulvenna, M., S. Anand, et al. (2000). "Personalization on the Net using Web Mining." Communications of the ACM **43**(8).

OKS Group (2003), "Web Mining", www.oksgroup.com

Quinlan, J (1992), "C4.5: *Programs for Machine Learning*, San Francisco, Calif: Morhgan Kaufmann.

QL2 Software (2003). "Bringing Order to data chaos: Improving productivity with information extraction tools". www.QL2.com.

Roussinov, D. and J. L. Zhao (2003). "Automatic discovery of similarity relationships through Web mining." Decision Support Systems **35**: pp.149-166.

Saunders, R (2004), "Genetic Patents and the Limits of IT", CSC World Dec 2003 – February 2004. (www.csc.com)

Thelwall, M. (2001). "Web log file analysis: backlinks and queries." Aslib Proceedings **53**(6).

Wang, S. (2002). "Knowledge maps for managing Web-based business." Industrial Management & Data Systems **102**(7): pp.357-364.

Wuthrich, B., D. Permunetilleke, et al. (1998). Daily Prediction of Major Stock Indices from textual WWW Data. 4th International Conference on Knowledge Discovery and Data Mining, pp.364-368

Yang, C., H. Chen, et al. (2003). "Visualisation of large category map for Internet browsing." Decision Support Systems **35**: pp. 89-102.