Position Paper: Concept Classification for Decentralised Search

Alan Dix, Filipos Karapetis, Genovefa Kefalidou
Computing Department, InfoLab21
Lancaster University
Lancaster, UK
+44 1524 510319

alan@hcibook.com

ABSTRACT

Whilst the Internet and the web use decentralised models, web search is currently highly centralised. We show how automatic concept classification enables a web search architecture that can be widely decentralized. In particular, the classifier presented borrows techniques from information retrieval in order to use the open directory project data set to classify into DMOZ categories. This allows indices to be divided based on conceptual categories and also enables the incorporation of hidden-web resources in a unified framework. In addition, we explore some human and economic issues that would allow or prevent the growth of true distributed web indices.

Categories and Subject Descriptors

C.2.4 [Computer-Communication Networks]: Distributed Systems – *distributed applications*. H.3.3 [Information Storage And Retrieval]: Information Search And Retrieval – *clustering, search process*. H.3.5 [Information Storage And Retrieval]: Online Information Services – *web-based services*. I.2.6 [**Artificial Intelligence**]: Learning – *concept learning*. I.7.2 [**Document And Text Processing**]: Document Preparation – *index generation*. K.4.1 [**Computers And Society**]: Public Policy Issues.

General Terms

Algorithms, Performance, Reliability, Human Factors

Keywords

Distributed search, automatic classification, hidden/invisible web

1. INTRODUCTION

The protocols of the Internet and the Web are open and free; the hardware is distributed and owned by multiple public and private institutions, the naming and other central features are administered by multiple institutions and ultimately regulated by the UN. This decentralisation was originally designed in order to avoid damage in nuclear war, but has been the key feature that has enabled its growth as global infrastructure. Similarly the open standards of the Web have allowed it to grow and flourish.

However, the Internet is not just infrastructure but also services, information and, perhaps most critically, the ability to find these. In contrast to the highly decentralised and open nature of the communications infrastructure, Web search is largely in the hands of a few large companies, notably Google, currently the market leader, but recently challenged by Microsoft who have now indexed more pages and sometimes vie with Google as the most frequent crawler in web logs.

This anomaly between the open and decentralised philosophy of the web compared corporate and centralised search has not gone unnoticed and there have been several projects aimed at various forms of distributed, decentralised and open-source web search. However, whilst the Open Directory Project (www.dmoz.org, resourced by Netscape) has made an effective 'community' alternative to Yahoo! and other bespoke directories, there has so far not been a similar success in broad web search.

This paper aims to address some of the key problems that have so far hampered the adoption of open search models. We describe how techniques of automatic content classification can be used to enable decentralised web search and the incorporation of hidden web resources (a.k.a. invisible/deep web).

The work is organised around a central vision of open web search not controlled by any single body, which would enable new research, reduce barriers to entry for innovative commercial enterprise, and reduce the fragility of a global search infrastructure based around a small number of companies and data centres. However, the detailed understanding, practical algorithms and tools that contribute towards this vision are also be of value to more dedicated repositories, such as digital libraries.

In some ways this research runs counter to the developments of the Semantic Web, which emphasises explicit meta-content markup. In contrast, our emphasis is on relatively unstructured sources and implicit semantics. However, these two are complementary approaches as automatic classification effectively adds a level of inferred semantics, which can be used to integrate with more structured data stores. For example, the SCORE systems uses document classification to disambiguate terms during metadata extraction [12].

2. AVOIDING DISTRIBUTED JOIN

One of the key technical problems for distributed search is that whilst crawling and to some extent indexing are relatively easy to distribute, the main cost of web search is in the actual processing of user queries. If the indices are spread over many servers how do queries get distributed and results gathered without massive network costs?

The simplest approach, is to split indices alphabetically (or based on hash) leading to small numbers of index servers being hit, one per search term. However, the returned result set, even as a collection of unique page id, would typically be enormous. In a

WWW 2006, May 22–26, 2006, Edinburgh, UK.

word frequency analysis of page titles and descriptions in ODP (see Fig 1), the words ranked 1000 in terms of frequency (e.g. forest, bulletin) occurred in around 1 in 2000 pages, even those ranked 10,000 (e.g. arrowhead, priory, backstreet) have frequency of about 1 in 10,000. So indices of 10 billion pages would return result sets of at least a million page ids on the majority of search terms. Of course for single-word searches this is not a problem results can be returned most relevant first; but for multi-word searches the results from several index servers would need to be merged and re-ranked – a distributed join problem, that appears at first sight to be totally infeasible or at least require radical solutions. Not surprisingly, NUTCH, the open-source search-engine project, dismiss distributed search processing as impractical

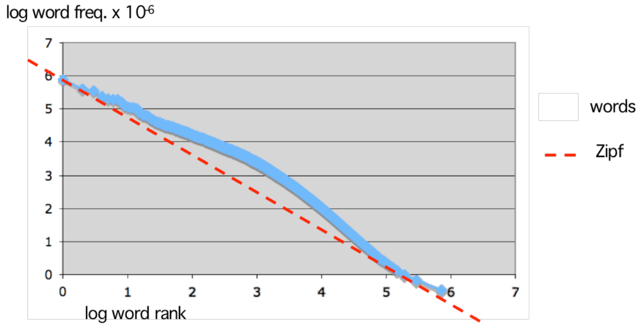

Fig 1. Word frequencies in ODP titles and descriptions vs. Zipf distribution (722,040 words, 3,002,045 pages)

Distributed join can be avoided if individual index servers are responsible for complete inverted indices of some portion of the space of web pages. For example, if an index is responsible for all pages within some set of domains, then each index server can perform the intersection of results for different search terms and produce a single ranking of the pages under its control. Of course now the problem is switched around – instead of a large distributed join of large result sets from a small number of servers, here we have a large union of small result sets from a vast number of servers. Each query needs to go to every server and potentially (and in even moderately common words this is likely) nearly every server may return results.

So our aim target is to somehow manage distribution in such a way as to avoid both large numbers of index servers being involved in every search and also to avoid large return sets requiring distributed join!

Happily, simple but powerful automatic classification techniques can help cut this Gordian Knot of distributed search. Assume we have index servers dedicated to complete page indices for dedicated topics (such as ‘pets’). Assume too that we are able to automatically classify search terms into areas such as ‘pets’. The search term can then be directed to a server specializing in pets. Even if the search term is ambiguous it can be directed to the most likely servers leading to a manageable distributed union. In order to assign pages to classified servers the same mechanism can be used during spidering. As pages are scanned they can be automatically categorised and allocated to appropriate index servers.

We have applied exactly this technique using an automatic classifier based on the DMOZ classification scheme. The classifier is automatic both in that it is automatically trained using the web pages in the ODP and is automatic in its classification of unseen resources and terms.

3. RELATED WORK

3.1 Distributed and federated search

In the traditional information retrieval literature there has been ongoing work on distributed indexing and search over many years including semantic partitioning [2, 5]. For example, the pSpace system uses term frequency vectors and maps regions of the high (300+) dimensional space to different servers [15]. Typically these systems involve closer coupling of servers than would be envisaged from a globally decentralised search infrastructure, but forms an important base point. Similarly work on federated databases whilst again operating in more 'controlled' environments have important lessons.

Turning to the Web, sporadic work on distributed search has been around for almost as long as search engines. Faced with the exponential growth in the number of web pages, commercial search engines looked towards distributed solutions, for example, Infoseek patented aspects of distributed crawling in 1997 and even the first academic versions of Google used distributed crawlers, even if within a closely coupled environment [4].

More recently, LookSmart have been using a downloadable SETI-like crawler, Grub, for collating basic change data (www.grub.org). However, crawling is only a small part of search-engine load and the more resource-intensive index serving is centralised in major engines. Whilst the enormous size of the indices necessitates some distribution this is within data centers with fast backbones, not over the Internet.

For digital libraries information sharing has long been important and standards for the interchange of metadata, search requests and results are mature. The pre-web Z39-50 standard dates back to 1995 as a formal standard and work on web standards include STARTS [8] and the Z39-50 'Zing' working group is developing the XML-based SRW protocol and CQL query language [17]. Bespoke services, such as Google, of course have their own XML APIs!

The open source Harvest project [3] (after a period of inactivity) is producing crawling and searching tools using a distributed architecture of Gathers (which crawl) and Brokers (which index and serve). This architecture supports federation but does not appear to be designed with large-scale distribution in mind. Harvest can index full text but it is optimised for sharing meta-information (author, title, etc.) automatically extracted from different file types.

Various digital library and resource sharing networks also tend to work at the level of sharing small amounts of meta-information for example the Metadata server at SUB Göttingen and the international PhysNet which uses Harvest technology. Metadata sharing is naturally at the heart of semantic web initiatives for distributed repositories such as Edutella [9].

3.2 Open source and open architecture tools

As well as numerous commercial web search engines there are also many open-source crawling, indexing and searching applications. Most of these are more suited for internal indexing of sites, but some, such as the Harvest project mentioned previously, are designed with larger scale use in mind. Even where the resources are distributed, many open search projects, for example ODISSEA, assume some form of centralised index of peer-peer shared resources [14]. The closest web-based base project to our work is the Java based Nutch project, which is currently hosted at the Internet Archive (www.archive.org).

Interestingly, as quoted earlier, the Nutch pages currently regard distributing indices as impractical – we intend to prove them wrong!

There are a number of digital library projects, some already noted above, but in particular the Greenstone Toolkit developed at Waikato. (www.greenstone.org), is well suited for indexing and storage within index servers.

3.3 Semantic inference and content integration

Semantic inference takes various forms including syntactic rules, natural language processing and latent semantic terchniques. The syntactic approach was used in CyberDesk [16], aQtive onCue [6] and Apple Data Detectors (developer.apple.com/sdk/) This uses templates, keywords, regular expressions, or hand-coded heuristics to identify possible types of data; for example "John Smith" is alphabetic, small number of words and has initial capitals so may be a name. Simpler rules are used by emailers and word-processors to identify email addresses and URLs in text. Natural language techniques have also been used, some based purely on grammatic forms, others using large tagged dictionaries such as WordNet [7] or combinations of the two. Latent semantic techniques are used widely for data mining and visualization, and also in web-based 'see also' services such as Alexa and are combined with the implicit structural semantics of links in Google. In traditional information retrieval systems they have also been used in choosing query servers for distributed repositories.

Semantic inference can be used during user query processing (as in onCue or various recent additions to Google) to select appropriate resources, or during data gathering to infer classifications and relationships or even complete ontologies. Examples of the latter include NLOS, which suggests potential key terms and relationships during requirements elicitation [10] and the SAI project focused on airport security scanning passenger lists for unusual relationships and potential threats [13].

Automatic classification systems may require extensive hand coding of training corpora or customisation. For example, the GRACE IST project uses ontology inference over free text documents, but (quoting the project web site www.grace-ist.org):

> *"Based on the hands on experience, it is estimated that integration of a small size ontology containing several hundred concepts requires approximately 10 workdays. The close assistance of the domain experts during this period is hereby stipulated."*

In contrast the techniques used in our work can leverage existing classified corpora, in particular the ODP data set, in order to classify vast ontologies. Clearly the results of automatic classifier training will not be the same as a hand-created rule set (although may not necessarily be uniformly worse), but are at least feasible in what would otherwise be intractable domains.

4. CONCEPT CLASSIFICATION

4.1 Background

The classification techniques we use were first developed by one of the authors when he was a director of a dotcom company during the late '90s. They were driven by an initial customer problem: suppose a user types "Chihuahua" into a search box in a eShopping directory, we would like pet shops to be returned even though the shops do not list every breed of dog, cat, budgerigar and goldfish in their keywords. To solve this problem we developed an automatic DMOZ classifier and also hand-classified the shops into the same scheme. When the user entered 'chihuahua', the system automatically recognised this as being connected with the Pets/Dogs category and also Travel/Mexico and also Taco Bell! The system therefore returned pet shops, travel agents and Taco Bell outlets (see Fig. 2) … the last of these did puzzle us until we realised that the Taco Bell mascot was a Chihuahua! A query for 'chihuahua poodle' returns Pets/Dogs with greater certainty.

As well as efficiently and simply solving the pet shop problem (and incidentally also suggesting travel agents), it was found that the techniques allowed improved web searching within the pages referenced by ODP page and also that they allowed the concept-clustered presentation of results described later.

In another demonstrator images, each with only small number of keywords, were automatically classified by putting the keywords into the classifier used for search terms. This allowed effective textual searching of the image database for queries using terms that did not appear in the images own keyword lists.

This last technique is the precisely what are applying for arbitrary spidered web pages.

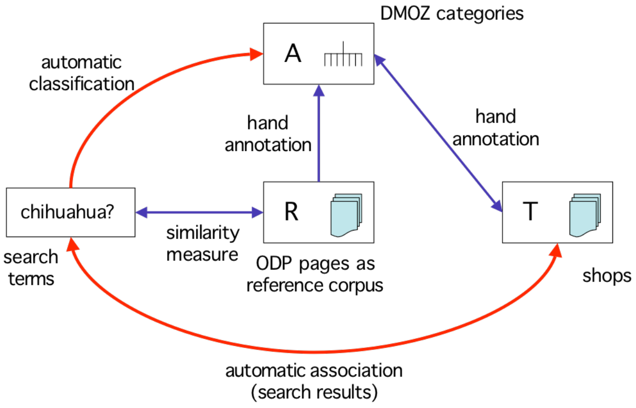

Fig 2. Using automatic query classification to find shops

4.2 How it works

Our concept classifier uses the fact that there is an existing hand-classified corpora in the Open Directory Project. At its simplest we are taking the words or phrases in a user query (or other term to be classified) and looking at which ODP pages contain these words. In the case of 'chihuahua' most of these pages would exist within the Pets/Dogs/Chihuahua[1] category and in Travel/Mexico/ sub-categories (but no longer Taco Bell as the mascot was withdrawn in July 2000). We can therefore infer that the word is in some way associated with these categories. The actual algorithms used are a little more subtle than this, but effectively leverage this broad method.

In the case of less precise terms, for example 'puppy, we would see occurrences of the word in more categories. If a term occurs

[1] To avoid full DMOZ names, shortened category names are used in several places. This also emphasises that the particular use of DMOZ categories is not critical, just the existence of some suitable category structure and classified corpus.

frequently in several sub-categories of a category then the category gets 'credit' for the word: a form of upward spreading activation. So the Pets/Dogs category will get high activation for 'puppy' whether or not it has many pages mentioning puppy directly classified to it.

On the other hand if a word occurs frequently in most sub-categories of a category, then it is less surprising that it is in any particular category. So there is a level of downward inhibition where the strength of association of 'puppy' with the Pets/Dogs/Chihuahua category is reduced because it is common to all of Pets/Dogs sub-categories.

Classification of multi-word terms involves a few more heuristics. Take the case of 'chihuahua poodle'. The word 'chihuahua' leads to a high level of activation of Pets/Dogs/Chihuahua and Travel/Mexico. The word 'poodle' gives rise to activation of 'Pets/Dogs/Chihuahua'. Between these the higher category 'Pets/Dogs' gets activated through upward spreading activation and the two breed sub-categories also have some increased activation due to downward spreading to each of their compliments from Pets/Dogs.

Although we have described the heuristics above in terms of a single level of activation, in fact we have needed several 'flavours' of activation. There is the raw activation level for a single word that only spreads upwards significantly to a super-category if a high proportion of sub-categories have high-activation and can involve downward *inhibition*. On the other hand, there is a more liberal spreading upwards and downwards related to a word for combining with other words.

Choosing the right levels and combinations of weightings is an area we are still experimenting with and have used different combination functions than in the original implementation. To date we have only applied variations of the algorithms with a single upward–downward pass for single word and then for combinations.

The Appendix shows examples of applying our experimental classifier to the terms 'chihuahua', 'chihuahua puppy' and 'chihuahua poodle' demonstrating several of the points above.. This is precisely the service used for search term classification in our systems. Web page classification uses a different service as it is more computationally expensive.

4.3 Pragmatics

Although we have described this process of activation spreading as if it were occurring at the moment of classification, in practice we pre-calculate the activation pattern for the most frequent 10,000 words and store the top 100 categories for each. This is the heaviest computational part of the process and the multi-word stage is then tractable.

We do not as yet index multi-word phrases such as 'great dane'. This would make it possible to efficiently weight categories referring to Great Dane dogs rank more highly than those about the Dane Threlked the Great.

In building indices, we have so far also only used the titles and descriptions in the ODP data dump and not spidered the actual pages referred to by the directory. This appears to give sufficient accuracy and it may be that these edited descriptions are more useful than the pages themselves.

In fact, we have noted that the word frequencies in these titles and descriptions do not follow a standard Zipf curve, but are slightly 'middle heavy' (see Fig. 1). It appears that editors use more middle-frequency words than one would expect to see in normal language. This makes some sense as they are trying to describe the pages as precisely as possible (hence pressure to use less common words), but still comprehensible (hence pressure not to use very uncommon words). This seems to match well the pattern of search-term choice for exactly the same reasons.

5. FROM CLASSIFICATION TO SEARCH

5.1 Search Architecture

The construction of a decentralized search using this classifier is now straightforward. Figure 3 shows the main components.

On the left are the web crawlers building the index. Pages are spidered using conventional techniques. These pages are then passed to the concept categorizer, which allocates a number of DMOZ categories with weightings. The page is then passed to the index servers corresponding to the highest weighted categories. Each server maintains a complete inverted index of all pages allocated to it using conventional techniques (or bespoke methods particular to the topic).

During search a broker agent is used as in other meta-search services. The search broker manages the user interface and passes the user's search to the concept categorizer. The search term is then passed to the index servers for the most relevant categories. Each index server returns a ranked list of search results and the broker returns the merged results to the user.

We have so far applied this process at a very small scale running the crawler only for experimental purposes. In order to populate the index servers for testing we have used the pages pre-allocated in the ODP. This has allowed us to test the search side independent of the spidering.

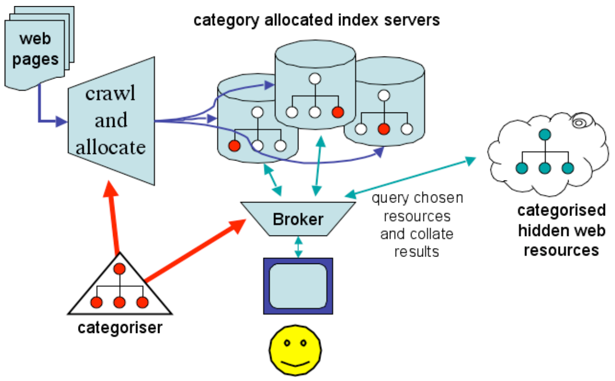
web pages
category allocated index servers
crawl and allocate
query chosen resources and collate results
categorised hidden web resources
Broker
categoriser

Fig 3. Search Architecture

5.2 Accuracy

We are currently happier with the classification of search terms than with the web page classification. The pages that cause problems are those that are home or welcome pages to sites as these tend to have relatively little text compared with 'decoration' such as menus. Arguably this does not matter if the more information-rich pages are well classified, but we would like to improve the classification of these entry pages.

One method that we hope will improve the classification of welcome pages is to use the classification of linked pages as an additional heuristic. If all the pages that are linked from a page concern a topic it is likely that the page also concerns the topic. A breadth-first crawler will be able to calculate this easily.

As we are still tuning these algorithms we have not as yet performed systematic accuracy measurements. The ODP dataset provides us with a useful ground truth for this as we can take hand-classified ODP pages and verify whether the automatic categorization matches the hand-classified categories.

However, it is perhaps not essential that the categorization is entirely 'accurate' in the sense that a human would agree with the categorization. Most crucial is that the categorization of search terms takes them to the index server where relevant pages are to found. If all pages concerning Chihuahua dogs were misclassified to Chihuahua in Mexico this would not matter so long as searches looking for Chihuahuas were similarly classified.

5.3 Harnessing the Hidden Web

One of the interesting aspects of concept-based query management is that accessing the hidden web (a.k.a. invisible web, deep web) virtually 'for free'. The Hidden Web refers to resources in publicly accessible databases and services that are available via the web, but are not web crawlable. They are estimated to include perhaps 10 times as much information as the visible web [11]. Commercial search engines are beginning to see the potential for incorporating this type of content and are making strategic alliances. For example, Yahoo!'s content acquisition programme invites public data sources to integrate their content into its service. Currently this is a major task for each data source and, presumably for that reason, appears to be limited to large repositories.

Automatic classification offers a particularly easy way to integrate hidden-web resources into 'normal' search. If the resources are classified by their DMOZ category, then the broker can pass relevant search terms to those hidden-web resources that belong to the categories inferred for the user's search. If the user type 'chihuahua' and there is a specialist dog database then a URL can be created on the results page that takes the user to the resource results in a single click.

Alternatively if the resource supports XML or other machine parsable results, such as the A9 OpenSearch standard (opensearch.a9.com), then the resource results can be integrated directly with web results in the same way as some search engines incorporate particular major resources, such as Wikipedia, today.

Whilst early experiments with this are promising, one problem that arises with resources not deliberately designed for such searching is the way they deal with general 'noise' words. For example searching the Internet Movie DataBase (www.imbd.com) for "films starring Julia Roberts" yields no results. The words "films starring" are good to tell you that IMDB is a good place to search, but the search engine would really like a name or film title to work with. This problem is even more acute for sites such as hotel finders where the query "hotels near LA1 4YR" would be no good at all if the query expected were a simple UK postcode.

To deal with this we are planning to combine the automatic classification to tell us what the query term is *about*, with techniques similar to those used in onCue to extract particular kinds of data (names, Post Codes, telephone numbers) from the query text. A hotel hidden-web resource can then be classified as being about Recreation/Travel/Lodging but requiring a string of the type "UK PostCode".

5.4 Sizing and Scaling

In this work we are thinking of index distribution over resources donated by institutions such as universities or small companies, not massive distribution onto individual desktop PCs. The latter are potential spiders within our architecture, as in Grub, but not for index servers in our scheme.

Basically as the number of index servers increase the accuracy required increases as the search terms and pages have to allocated to very precise categories. Larger donated resources can store a larger slice of the page space and thus can be allocated categories 'high' enough to be accurately classified.

The number of severs required for a single replica can be estimated easily. We assume indexing of 10 billion pages (Google have stopped quoting a figure, but it was creeping towards this), with approx 10Kbytes for an inverted index per page and up to 5 categories allocated per page. This means 50 billion page–category instances so around 0.5 petabyte total storage. If participants allocated 0.5 terrabytes 1000 particpants would be required for a single replica. Furthermore this would mean breaking the concept space into approximately 2000 parts (for reasons described later a site would normally server at least two category slices). Given a typical DMOZ branching factor of 20–30, this means slicing at level 2 or 3 in the DMOZ category structure – high enough to be easily categorized.

The fact that a page is allocated to a relatively small number of categories is very important. To see why, assume instead a word-based hierarchy were used to allocate index servers: there would be perhaps 500 word instances per page and so 5 trillion page/word instances. However, lexical indexing would be totally accurate and hence allow finer granularity, so lexical techniques would be more appropriate for PC-level distribution, albeit with an order of magnitude more contributors required for a basic distribution.

To realize our scheme in practice would require some form of decentralised registration system so that brokers can know where index servers are for various categories. However, dividing the category space into only a few thousand categories means that brokers can maintain complete caches of index server addresses.

The classifier itself currently uses approximately 2Gbytes of disk space, so would be able to be easily replicated on brokers and perhaps even crawlers. However, for web-page allocation a web service approach may be preferable. This is because it is sufficient to pass the 50 or so least-common words from the web page to the classifier rather than the full text. Using a web service would make possible micro-crawling, such as browser plug-ins that scan visited pages only without having to replicate the classifier to each spider.

5.5 Bootstraping

We will deal later with why institutions might choose to offer resources to such a scheme, however undoubtedly one could not immediately sign up 1000 volunteer institutions! A key issue with the uptake of any technology is whether there is a credible path from no use to widespread use.

Happily the use of concept-based indices makes it easier to start small with niche indices gradually growing to cover wider areas. Niche indices can combine hidden web resources, bespoke data and also crawling seeded from known high-quality sites. The concept classification of search terms means the search broker can know when a term can be looked for in one of the existing indices and when the user should be forwarded to a more general resource.

The other obvious means to bootstrap this process is through the existing classified Open Directory Pages. These can be used in two ways. First, as they are pre-classified they can be searched as is. This has been used in our experiments so that we can have a 'complete' search without any crawling at all! Of course, this is limited to the 3 million ODP classified pages, but these are at least hand selected for relevance. In addition, however, these give a set of pages with some level of quality that form an obvious seed for crawling covering a wide range of areas. Assuming that the initial pages are of good quality then this is likely to mean that other high-quality pages are spidered sooner.

6. HUMAN ISSUES

6.1 About vs. Containing Searches.

Concept classification is very good at telling you where to find pages *about* things like "Chihuahua". However, it is less good at answering the query "I know I saw a page about something other than dogs that mentioned the word Chihuahua". That is looking for pages *containing* the word "Chihuahua". This is the opposite of standard web search engines, which start off with the pages containing a term and then use ranking and possibly clustering to try, in a way, to recover meaning.

Of course, if the term is very common then no search engine would be helpful, however if the user recalls that the page was something to do with the Eiffel Tower then a search for "Eiffel Tower Chihuahua" would be expected. This effectively means: Look for pages *about* "Eiffel Tower" containing the word "Chihuahua" – but of course this is implicit not explicit in the query.

To some extent this implicit intention can be recovered automatically from the query. The query "Eiffel Tower Chihuahua" gets allocated to the categories Regional/Europe/France and also categories to do with Pets/Dogs. However, in the classification to Regional/Europe/France only the term "Eiffel Tower" will have been important, so it is clear that "Chihuahua" is the word to be looked for in these pages, but equally "Chihuahua" was important for Pets/Dogs so the term "Eiffel Tower" will be looked for in dog pages. However, it cannot tell whether the user means pages about Chihuahuas containing the words "Eiffel Tower" as compared with pages about the Eiffel Tower containing the word "Chihuahua".

This suggests that users could be offered ways to make this explicit and hence allow more precise queries. This would be particularly powerful when trying to find a page about some common term that is used in a specific way within a discipline, or a person with a common name in a particular field. However, studies of actual search queries tend to show very little use of advanced search facilities amongst even expert users [1], so the utility of this may be limited for normal searches. the exception might be for search boxes on subject-specific portals or sites. For example, a site about Chuhahuas might have a web search where there is a hidden added term "about:Chihuahua", rather like Google site-search boxes effectively add "site:dogs.org".

6.2 Classification of Results

Concepts can also be used to group results. For example a search for "Chihuahua" on Google yields a mix of dog sites and those for Chihuahua the place in Mexico. In this case, both appear mixed even on the first results page. In other cases the most popular use of a term may crowd out the use that you are after (no Taco Bell related Chihuahua pages!).

Because the pages are classified in our scheme, it is possible to know that these pages cover a number of distinct areas and to group results accordingly. Instead of a plain list of results, concept-grouped results mean that a small number of canine, geographical sites and those about the Taco Bell's mascot can all be presented on page 1 with 'more like these' links. Similar techniques are offered using post hoc clustering on sites such as clusty.com and wisenot.com.

6.3 Adoption and Economics

Whilst we have mainly discussed the technical issues related to concept classification for decentralised search, there is perhaps a far more critical question that is common to all peer or volunteer schemes: why would anyone donate resources to it? In fact, the experience of the internet is that people, public institutions and commercial companies do donate significant resources to projects for 'public good' including software mirrors, free data resources, etc. Often these serve as publicity or, where there is a 'front end' as a source of advertising revenue. However, there are a growing number of RSS and XML feeds where the backend provider gets little apart from the knowledge that they are helping the broad web community.

However, the use of classified indices also offers additional benefits to service providers. By agreeing to provide, for example, an index for the Pets/Dogs category the donor is being given access to the repository of pages that can be mined in ways other than standard web search. This may be used to provide value-added services within the category or for internal use. For example, there is a growing market for web intelligence, companies who use web crawls to watch competitors or observe market conditions (e.g. ultimathule.net). These could benefit by having broader repositories than they could gather themselves.

There will of course be categories that would be oversubscribed (such as university computing departments hosting computing categories, or companies hosting stock-market related information), but other categories (dogs?) may have no takers! One of the reasons for suggesting that index servers should serve at least two distinct category slices is so that part of the 'price' of choosing a category of value to you is that you must also serve a less popular category chosen for you.

For the front-end web brokers the value is much more apparent as they are in a position to reap advertising and eCommerce revenue like standard search companies. In order to differentiate themselves they may chose to subscribe to value-added services from index servers perhaps adding bespoke data to crawled pages, or using topic-specific ranking algorithms, thus creating a 'backend' market.

It is important to note that an open search architecture does not preclude commercial use, but in fact potentially makes it a more open market. This is similar to the way in which the Open Directory Project is not a competitor to commercial directories, but complements commercial web services. For example, Google uses ODP for its directory and the reason Netscape hosts and promotes ODP is partly to use it in its own portal. Similarly an open web-search infrastructure makes it easy for new enterprises to enter the market based on improved interfaces or ranking algorithms, or to include specialised niche searches.

6.4 Cheats and Liars!

One of the key arguments for open source by the NUTCH project is the need to trust that results of searches are ordered by relevance to the users not (unless clearly signalled) based on who pays for inclusion. Distributed search leads to different issues of trust; for example, how can one know that a spider is not introducing false keywords for pages? Centralised search engines have had to cope with attempts using invisible text or similar techniques to subvert their indexing algorithms. Similarly the results of decentralised servers may need to be verified and there may need to be separate validation services perhaps cross-matching results from multiple index servers, verifying classifications, etc. Again this has the potential to be a commercially valuable service.

7. FUTURE AND RELATED WORK

We have already noted various areas for further work, notably in optimizing parameters in concept classification, dealing with text-sparse page, techniques for dealing with rare, non-domain specific words (e.g. personal names), and managing more fragile hidden-web resources. These are already on our own short and medium term agenda.

In addition there are a whole set of issues that would need to be solved in order to move from small-scale prototype to large-scale deployment. Some of these are standard issues such as managing replication between index servers for the same concept areas. Some are common to any scheme for decentralised search, such as management, formal agreements and the issues of misuse. Some are more specific to our technique such as dealing with category evolution or load balancing when world events make a particular topic 'hot'.

As well as applications within the web-domain, we are aiming to adapt the concept classification techniques to smaller personal ontologies as part of work on personal information management and task-based interaction.

In the near future we also aim to make the laboratory prototype available on the web focused on a niche area with 'backup' general web search (but most likely human–computer interaction rather than dogs!). The concept classifier is available (for small numbers of words, not page classification) at the address in the appendices for light use by other researchers.

Whilst many detailed issues remain to be resolved, we believe the proposals in this paper remove a major stumbling block from decentralised search and offer an achievable path for future deployment.

8. ACKNOWLEDGMENTS

This work was supported by the TIM project part of the EC funded DELOS Network and by the EPSRC funded EQUATOR Project. Thanks also to the aQtive team now in diaspora.

9. REFERENCES

[1] Aula, A. (2003) Query Formulation in Web Information Search. In Isaías, P. & Karmakar, N. (Eds.) Proceedings of IADIS International Conference WWW/Internet 2003, Volume I, pp. 403-410. IADIS Press.

[2] R. Baeza-Yates & B. Ribeiro-Neto. Modern Information Retrieval.. Addison Wesley. 1999.

[3] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber & M. F. Schwartz. The Harvest Information Discovery and Access System. Computer Networks and ISDN Systems 28 (1995) pp. 119-125.

[4] S. Brin & L. Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7 / Computer Networks 30(1-7): 107-117 (1998)

[5] J. Callan. Distributed Information Retrieval. In W.B. Croft, ed.. Advances in Information Retrieval, Chapter 5, pages 127–150. Kluwer Academic, 2000.

[6] A. Dix (1999). Design of User Interfaces for the Web (invited paper). In User Interfaces to Data Intensive Systems, UIDIS. pp. 2-11 IEEE Computer Society, 1999.

[7] C. Fellbaum (ed.). WordNet: an electronic lexical database. MIT Press, 1998.

[8] L. Gravano, C. K. Chang, H. Garca-Molina & A. Paepcke. STARTS: Stanford proposal for Internet meta-searching. In Proceedings of the 1997 ACM SIGMOD Conference, 1997.

[9] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palmer & T. Risch. EDUTELLA: A P2P networking infrastructure based on RDF. November 2001. http://edutella.jxta.org/reports/

[10] P. Sawyer & K. Cosh. Supporting MEASUR-driven analysis using NLP tools". In Proc. 10th International Workshop on Requirements Engineering,: Foundations of Software Quality (REFSQ'04), Riga, Latvia, June 2004.

[11] C. Sherman & G. Price. The Invisible Web: Finding Hidden Internet Resources Search Engines Can't See. CyberAge Books, 2001

[12] A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut & Y. Warke.. Computing, July-August 2002, pp. 80–87.

[13] A. Sheth, B. Aleman-Meza, I. B. Arpinar, C. Bertram, Y. Warke, C. Ramakrishnan, C. Halaschek, K. Anyanwu, D. Avant, F. S. Arpinar & K. Kochut. Semantic Association Identification and Knowledge Discovery for National Security Applications. Journal of Database Management. Special Issue of on Database Technology for Enhancing National Security, L. Zhou. (ed.) 2005 (in press).

[14] T. Suel, C. Mathur, J. Wu, J. Zhang, A. Delis, M. Kharrazi ,X. Long, & K. Shanmugasundaram . ODISSEA: A Peer-to-Peer Architecture for Scalable Web Search and Information Retrieval. In International Workshop on the Web and Databases (WebDB). June 12–13, 2003, San Diego, California.

[15] C. Tang, S. Dwarkadas, and Z. Xu. On Scaling Latent Semantic Indexing for Large Peer-to-Peer Systems. In The 27th Annual International ACM SIGIR Conference (SIGIR'04, ACM Special Interest Group on Information Retrieval). Sheffield, UK. July, 2004.

[16] A. Wood, A. K. Dey & G. D. Abowd. CyberDesk: Automated Integration of Desktop and Network Services, Proceedings of CHI'97, ACM Press, pp. 552-553. 1997.

[17] ZING. Z39.50 International: Next Generation. Library of Congress Network Developmeent and MARC Stabdards Office. Nov. 2004. http://www.loc.gov/z3950/agency/zing/

Appendix – categorization examples

Example 1 – chihuahua

Classifying ‘chihuahua’ would use the URL:

http://www.meandeviation.com/odp2/ui/magic-marker-v1.php?search=chihuahua

This results in the following classes (with relevance in range0–100, only those with relevance ≥ 50 shown):

```
Recreation/Pets/Dogs/Breeds/Toy_Group/Chihuahua            100
Recreation/Pets/Dogs/Breeds/Toy_Group                       64
Regional/North_America/Mexico/States/Chihuahua              55
Recreation/Pets/Dogs/Breeds                                 54
Recreation/Pets/Dogs                                        54
Recreation/Pets                                             53
Recreation                                                  51
Shopping/Pets/Cats_and_Dogs/Clothing_and_Accessories        51
Regional/North_America/Mexico                               50
Regional/North_America/Mexico/Business_and_Economy          50
Shopping/Pets/Cats_and_Dogs                                 50
Shopping/Pets                                               50
Regional/North_America/Mexico/States                        50
News/Online_Archives/CNN.com/2003/August                    50
Regional/North_America/United_States/Alabama                50
Science/Social_Sciences                                     50
```

Note that the name of the category is NOT used in the classification

Note also that August 2003 is when news items discussed Taco Bell’s decision to retire its chihuahu mascot!

Example 2 – chihuahua puppy

Classifying ‘chihuahua puppy’ would use the URL:

http://www.meandeviation.com/odp2/ui/magic-marker-v1.php?search=chihuahua+puppy

This results in the following classes:

```
Recreation/Pets/Dogs                               100
Recreation/Pets/Dogs/Breeds                         99
Recreation/Pets                                     97
Recreation                                          92
Shopping/Pets                                       88
Recreation/Pets/Dogs/Breeds/Toy_Group/Chihuahua     87
Recreation/Pets/Dogs/Breeds/Toy_Group               56
```

Note how the generic category Recreation/Pets/Dogs/Breeds has risen to top place, also the Mexican state does not appear.

Example 3 – chihuahua poodle

Classifying ‘chihuahua poodle’ would use the URL:

http://www.meandeviation.com/odp2/ui/magic-marker-v1.php?search=chihuahua+poodle

This results in the following classes:

```
Recreation/Pets/Dogs/Breeds                                                       100
Recreation/Pets/Dogs                                                               99
Recreation/Pets                                                                    98
Recreation                                                                         95
Shopping/Pets                                                                      94
Recreation/Pets/Dogs/Breeds/Toy_Group/Chihuahua                                    93
Recreation/Pets/Dogs/Breeds/NonSporting-Utility_Group/Poodle/Clubs                 67
Recreation/Pets/Dogs/Breeds/NonSporting-Utility_Group/Poodle                       64
Recreation/Pets/Dogs/Breeds/Toy_Group                                              60
Recreation/Pets/Dogs/Breeds/NonSporting-Utility_Group/Poodle/Rescues_and_Shelters  59
Recreation/Pets/Dogs/Breeds/NonSporting-Utility_Group/Poodle/Pets                  59
Recreation/Pets/Dogs/Breeds/NonSporting-Utility_Group                              54
Regional/Europe/United_Kingdom/Recreation_and_Sports/Pets/Dogs/Breeds/Poodle       52
Regional/North_America/Mexico/States/Chihuahua                                     51
```

Note that again the generic category Recreation/Pets/Dogs/Breeds is in top place above the specific breed categories. This time the Mexican state does appear, but is low in the list.