2008 Second IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2008)

IR Issues for Digital Ecosystems Users

Dengya Zhu and Heinz Dreher

Curtin University of Technology, GPO Box U1987, 6845 Perth, Western Australia,
e-mail: dengya.zhu@postgrad.curtin.edu.au, h.dreher@curtin.edu.au,

Abstract—**the purpose of this research is to discuss some challenges of information retrieval, especially Web information retrieval, in digital ecosystems from a user's perspective. As a dominant search tool, search engines usually return millions of search results in a long flat list in which many or even most of the results can be irrelevant. The long flat list conveys nothing about knowledge structure related to the retrieved results and personal search preferences and interests are not explored. Although some search engines try to cluster the Web results, the automatically formed titles and knowledge hierarchy is prone to mismatching the searcher's human mental model. In digital ecosystems, while many different search tools are available, they are not integrated. To address these issues, a search framework which combines categorization, clustering, ontology, and personalization is proposed, and thus the quality of search results in digital ecosystems is expected to be boosted.**

Index Terms—**Web information retrieval, personalization, digital ecosystems, search engines, categorization, clustering.**

I. Introduction

Information retrieval on the WWW is far from perfect [16]. As a dominant search tool, search engines usually return millions of search results in a long flat list in which much of the returned results set can be irrelevant; and despite the long list of results, users' information needs are frequently not satisfied. The flat list of information items conveys nothing about the latent knowledge structure related to the retrieved results. To address this problem, some search engines try to cluster the returned results by grouping them into an automatically formed knowledge hierarchy [6][8]. However, the automatically formed knowledge hierarchy looks very strange from a human's perspective; it mismatches the human mental model [30]. Text categorization [5][24] , which employs a human edited knowledge structure, can organize Web search results into a human friendly form, however training data obtained from experts are expensive.

Search results personalization is considered a promising approach which addresses the adaptability of information retrieval system to the needs and interests of individual users [20], however, explicitly constructed user profiles suffer from the problems of extra user burden, inaccurate preferences description, and concept drift [7]. Implicitly constructed user profile using machine learning approaches also needs to deal with problems such as the need for large data set, the need for labelled data, concept drift, and computational complexity [27].

A Web navigator such as *Yahoo! Web Directory* [28] is used as a portal for Web information retrieval. *Yahoo! Web Directory* can provide relevant information under a specific topic, however, with the rapid growth of the Web, finding a "right" topic is becoming very difficult, and the *recall* of the search-results is very low despite the *precision* being high[1].

In digital ecosystems, search tools are still not integrated into a user friendly interface. Users can retrieve information from diverse information sources using different search tools. Learning how to use these different tools increases training cost.

To address some of the issues above, a search framework is proposed and combines the power of text categorization, clustering, ontology, and personalization in an effort to boost the quality of Web information retrieval.

The paper is organized as follows: in section II, issues related to Web information retrieval and in digital ecosystems are first presented from a user's perspective; the reasons behind these phenomena are then analyzed and discussed; section III proposes a search framework aimed at addressing some of the issues discussed; section **Error! Reference source not found.** suggests future work regarding this research; and finally section V, concludes the paper.

II. Problems of Information Retrieval

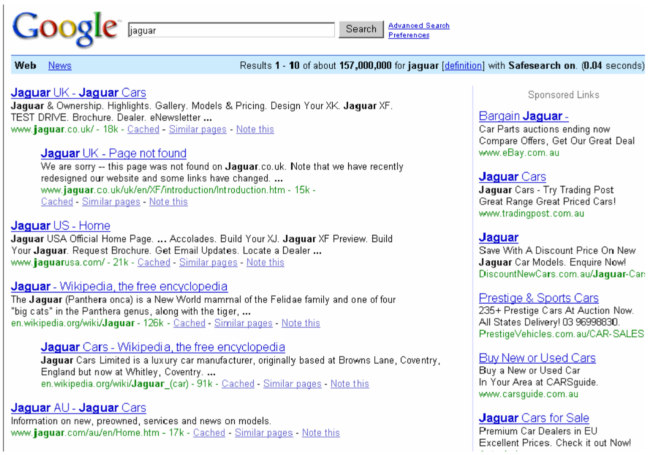
In this section, some issues of Web information retrieval are considered; reasons for some issues are analyzed and discussed.

A. Information Overload – 64 Million Search Results

Search engines usually return a long list of Web search results as shown in Fig. 1, where "jaguar" is used as a search-term. As can be seen from the figure, Google returned 157 million of search results, and within the first 10 returned results, only three of them are about the animal jaguar. Using "George Washington" as search-terms to retrieve information about the American boxing trainer, Google retrieved 39 million search results (retrieved on Nov 1st, 2007). However, among the top 100 retrieved results, not one is relevant [30]. In this so called "ill-defined queries" [18] case, there may be only a few low-ranked items returned and they are rarely noticed by most users.

The miscellaneous demography of Web users requires that Web service providers should not expect too much to their users. In fact, when a car salesperson is searching for information about the jaguar car, in the mind exists the concept "jaguar" - nothing but a jaguar car. Furthermore, although using an additional phrase such as "jaguar car" can make search engines return information items containing that text string, it may also cause search engines to

[1] Precision and recall are two measures of an information retrieval system. Recall measures how well a system retrieves only relevant documents; precision evaluates how well the retrieved documents are relevant.



Results **1** - **10** of about **157,000,000** for **jaguar** [definition] with **Safesearch on**. (**0.04** seconds)

Jaguar UK - **Jaguar** Cars
Jaguar & Ownership. Highlights. Gallery. Models & Pricing. Design Your XK. **Jaguar** XF. TEST DRIVE. Brochure. Dealer. eNewsletter **...**
www.**jaguar**.co.uk/ - 18k - Cached - Similar pages - Note this

Jaguar UK - Page not found
We are sorry -- this page was not found on **Jaguar**.co.uk. Note that we have recently redesigned our website and some links have changed. **...**
www.**jaguar**.co.uk/uk/en/XF/introduction/Introduction.htm - 15k - Cached - Similar pages - Note this

Jaguar US - Home
Jaguar USA Official Home Page. **...** Accolades. Build Your XJ. **Jaguar** XF Preview. Build Your **Jaguar**. Request Brochure. Get Email Updates. Locate a Dealer **...**
www.**jaguar**usa.com/ - 21k - Cached - Similar pages - Note this

Jaguar - Wikipedia, the free encyclopedia
The **Jaguar** (Panthera onca) is a New World mammal of the Felidae family and one of four "big cats" in the Panthera genus, along with the tiger, **...**
en.wikipedia.org/wiki/**Jaguar** - 126k - Cached - Similar pages - Note this

Jaguar Cars - Wikipedia, the free encyclopedia
Jaguar Cars Limited is a luxury car manufacturer, originally based at Browns Lane, Coventry, England but now at Whitley, Coventry. **...**
en.wikipedia.org/wiki/**Jaguar**_(car) - 91k - Cached - Similar pages - Note this

Jaguar AU - **Jaguar** Cars
Information on new, preowned, services and news on models.
www.**jaguar**.com/au/en/Home.htm - 17k - Cached - Similar pages - Note this

Sponsored Links

Bargain **Jaguar** -
Car Parts auctions ending now
Compare Offers, Get Our Great Deal
www.eBay.com.au

Jaguar Cars
Jaguar Cars - Try Trading Post
Great Range Great Priced Cars!
www.tradingpost.com.au

Jaguar
Save With A Discount Price On New
Jaguar Car Models. Enquire Now!
DiscountNewCars.com.au/**Jaguar**-Car

Prestige & Sports Cars
235+ Prestige Cars At Auction Now.
All States Delivery! 03 96998830.
PrestigeVehicles.com.au/CAR-SALES

Buy New or Used Cars
Buy a New or Used Car
In Your Area at CARSguide.
www.carsguide.com.au

Jaguar Cars for Sale
Premium Car Dealers in EU
Excellent Prices. Check it out Now!

Fig. 1 Search results of "jaguar" from Google (Nov 1st, 2007)

miss some relevant documents which do not contain "jaguar car". Research has shown that user prefer to use one word or very short phrase as search-terms [13]. Users have difficulty in expressing their information needs [2][3]. Therefore, Web search providers should adopt technical solutions to address the issue and not ask and expect users to submit well-defined, machine understandable search-terms.

Information Retrieval is different from database searching. Information retrieval "deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested." [2] Database searching, on the other hand, focuses on exact matching. Database retrieval languages (such as SQL) aim at searching all objects which match clearly defined conditions like those in an algebraic expression. Any mismatch among the thousands of objects is thus an error [2]. This approach is obviously not suitable for text document retrieval because if data searching techniques were used in information retrieval, the data retrieval systems would only return a list of documents if—and only if—they contain the exact search-term (query), and without concern about how to satisfy users' information needs. Information retrieval system is more tolerant to this kind of inaccurate matching error. This is mainly because an information retrieval system usually deals with natural language text which is not always well structured and could be semantically ambiguous [2].

Another difference between IR and data retrieval is that an information retrieval system always aims at returning information items which are relevant to the subject or topic to the user information needs as conveyed by search-terms. Relevance judgment, per se, is a subjective matter and a nontrivial issue [17]. To be effective in trying to satisfy users' information needs, an IR system must interpret the contents of the text objects in a collection, and rank the retrieved results according to their degree of relevance to the search-terms. This interpreting process involves extracting both syntactic and semantic characteristics from the text objects, and using the extracted information to match users' information needs. In an information retrieval system the concern is not only about the syntactic interpretation of search-terms and text objects, but also the relevance of an object to users' information needs [2].

One effective, efficient and widely accepted information retrieval model is Vector Space Model [2][23]. In this model, documents and search-terms are represented by a vector in a high dimensional vector space where each term appears in the documents or search-terms is corresponding a dimension in the vector space. The similarities between search-terms and documents are estimated by the cosine values between vectors represent the documents and the search-terms. The final search results are ranked according to the calculation and presented to users.

Issues arise for an information retrieval system when a search-term has more than one meaning (polysemy) and one meaning can be represented by more than one term (synonym). In the first scenario, if a document contains the search-term, no matter what the subject / topic the document is about, this document is regarded as relevant if the cosine similarity between the search-term and document is high. For example, if one Web page is about jaguar car and another is about animal jaguar, both Web pages will be in the search results if "jaguar" is used as search-term. Web search engines are also a kind of information retrieval system and will be affected by the polysemy and synonym issues, despite other techniques such as PageRank [19] and HITS (Hypertext Induced Topic Search) [15] being employed to improve the relevance of Web search results.

B. Mismatching Results – High Recall, Low Precision

As mentioned above, Google returns 157 million search items when "jaguar" is submitted as a search-term. In this case, the search results have a very high *recall*, because there are as many as 157 million items that contain the search-term "jaguar" However, if the needed information is about animal jaguar, from the user's perspective, the *precision* of the search results is very low, because among the answer set, most of the search results are irrelevant.

Low *precision* high *recall* problem may be caused by the following three reasons. The first reason is the inherent polysemy characteristic of natural languages. For example, because the search-term "jaguar" may refer to different things, it is very difficult, or impossible, for a search engine to return only relevant search results for an individual information seeker. One searcher may want to retrieve some information about the animal jaguar; another seeker may want to search where to buy a jaguar car. Without interaction with the individual user, it is unreasonable for a search engine to return only information concerned with a jaguar car, or only to return search results related to the animal jaguar. Therefore, for a search engine, there is no choice but to try its best to return all the documents that contain the term "jaguar". However, for a given user, the information about animal jaguar may be the only concern; thus from this user's perspective, the *precision* of the search results is very low, because many of the search results are not relevant to the user's information need.

Another reason is that the information retrieval model employed by most search engines is based on the term-weighting strategy [23] which relies on the idea of "taking words as they stand" and "counting their stances" [11]. This implies that search engines perform only syntactic comparison between search-terms and the indexed terms of document repositories [1][14]; semantic characteristics of the search-terms and the documents are ignored.

A third reason for the low *precision* high *recall* problem is users' search habits of using very short search-terms. Research [12][13] shows that more than 70 per cent of queries are composed of one, two, or three terms. However, in many situations, using more than one word as a search-term allows search engines to return more relevant search results than just using only a single word as a search-term. For example, when searching for information about the boxing trainer "George Washington", if the search-terms are "George Washington" + boxing, the returned search results are mainly related to the boxing trainer George Washington. Training users to select proper search-terms is out the scope of this research.

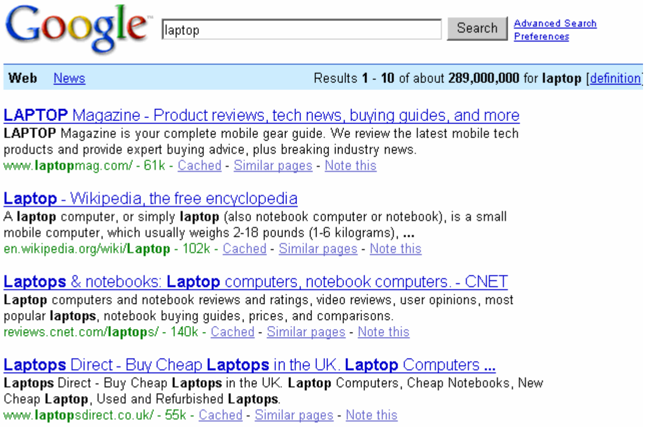
These three reasons are the sources of the high *recall*, low *precision* problem of search engines.

C. Missing Relevant Document – Low Recall

Despite of millions of research results being returned, the issue of missing relevant documents is still problematic for search engines in some circumstances. The problem of low *recall* of search results may be caused by both the polysemy and synonymy characteristics of natural languages but another, and more complex reason is that searching for information in one category usually would not retrieve information of its subcategories.

The polysemy feature of natural languages may cause relevant documents to be missed. As discussed in the previous section, when performing a search by using a search-term "George Washington" to retrieve information about the boxing trainer George Washington, most of the search results are about the first American president, George Washington, or George Washington University, or George Washington Bridge. Few search results are about the boxing trainer (not one returned result within the top 100 ranked documents relates to the boxing trainer). Interacting with the user [32] or the stored user-profile (refer to part F of this section) one may be able to determine a set of appropriate categories to focus the search and hence increase recall.

Search results returned by search engines are mainly decided by two calculated factors, one is a syntactic similarity comparison between documents (Web pages) indexed by search engines and the search-terms; another is the *authority* of the Web page [15]. The search results are then ranked according to the calculated similarities and their authorities. If the similarity and *authority* of a Web page is higher than a so-called *threshold*, the Web page is to be returned as a search result, otherwise, it will not be considered relevant to the search-term and will not appear in the list of search results. Compared with the Web pages about the first American president George Washington, George Washington University, or George Washington Bridge, the *authorities* of the pages about the boxing trainer George Washington are obviously very low, and even lower than the given threshold, and thus excluded from the returned list. In this circumstance, although there are millions of search results, very few of the relevant documents are presented in the search results set; the *recall* is thus very low.



Results **1** - **10** of about **289,000,000** for **laptop** [definition]

LAPTOP Magazine - Product reviews, tech news, buying guides, and more
LAPTOP Magazine is your complete mobile gear guide. We review the latest mobile tech products and provide expert buying advice, plus breaking industry news.
www.**laptop**mag.com/ - 61k - Cached - Similar pages - Note this

Laptop - Wikipedia, the free encyclopedia
A **laptop** computer, or simply **laptop** (also notebook computer or notebook), is a small mobile computer, which usually weighs 2-18 pounds (1-6 kilograms), **...**
en.wikipedia.org/wiki/**Laptop** - 102k - Cached - Similar pages - Note this

Laptops & notebooks: **Laptop** computers, notebook computers. - CNET
Laptop computers and notebook reviews and ratings, video reviews, user opinions, most popular **laptops**, notebook buying guides, prices, and comparisons.
reviews.cnet.com/**laptop**s/ - 140k - Cached - Similar pages - Note this

Laptops Direct - Buy Cheap **Laptops** in the UK. **Laptop** Computers **...**
Laptops Direct - Buy Cheap **Laptops** in the UK. **Laptop** Computers, Cheap Notebooks, New Cheap **Laptop**, Used and Refurbished **Laptops**.
www.**laptop**sdirect.co.uk/ - 55k - Cached - Similar pages - Note this

Fig. 2 Search results of "laptop" from Google (Nov 1st, 2007)

The inherent synonymy problem of natural languages is another reason for missing relevant search results. For example, "laptop" and "notebook" are synonymous; both refer to a very small portable computer that can be used on one's lap (notebook is usually believed smaller than a laptop). When searching for "laptop", Google returns 289 million search results, as shown in Fig 2. When searching for "notebook", Google returns 14 million search results (retrieved on Nov 1st, 2007). However, the search results of the two synonyms are different, although many of the returned Web pages of both search result sets are about "a portable, usu. battery-powered microcomputer small enough to rest on the user's lap" (http://dictionary.reference.com/search?r=8&q=laptop).

Using acronyms or abbreviations of some search-terms also causes search engines to return different search results. This problem is similar to that of synonymy. "AI" is an acronym of "Artificial Intelligence". Using "Artificial Intelligence" as a search-term, Google returns 33 million search results (retrieved on Nov 1st, 2007); when using "AI" as a search-term, there are 666 million returned items (retrieved on Nov 1st, 2007). Among the first 20 top ranked search results of the two search-terms, only eleven returned items are common to both.

Searching for information relevant to one category usually will not obtain documents relevant to its subcategories. For example, when searching for "machine learning", the search results will not contain "genetic programming", a subcategory of machine learning and artificial intelligence. This is because the retrieval model utilized by search engines only matches documents that are syntactically similar to search-terms. As a result, in the situation when searching for information about one category, search engines usually do not return information of its subcategories, and the *recall* of the search results is thus very low.

Using WordNet (http://wordnet.princeton.edu) or thesaurus can alleviate the problem caused by subcategory, and synonym characteristic of natural languages [22]. However, some synonymies may further introduce polysemy problems. In the example of "notebook" and "laptop" above, except for having a similar meaning to "laptop", the term "notebook" also represents a film named "notebook", an "electronic notebook of Google" and, of course, the commonsense notebook – a book of paper on which notes are written. Among the first 20 retrieved search results of "notebook", only five of them have a similar meaning to "laptop". Using "notebook" as a synonym may improve *recall* of search results; however, the *precision* of the search results may further be deteriorated.

D. Flat List vs. Structured List

As discussed above, most search engines arrange search results according to ranking algorithms that rank documents in higher priority according to the document's literal similarities to the given query [1]. Ranked documents are considered relevant to a user's query in descending order, that is, the first several documents are more relevant to the user's query than the rest of the search results. However, because all the returned search results are presented in a flat list, a user may have to check hundreds of Web snippets to pick up useful information. Finding a relevant document among the returned Web search results is like finding a needle in a haystack [2].

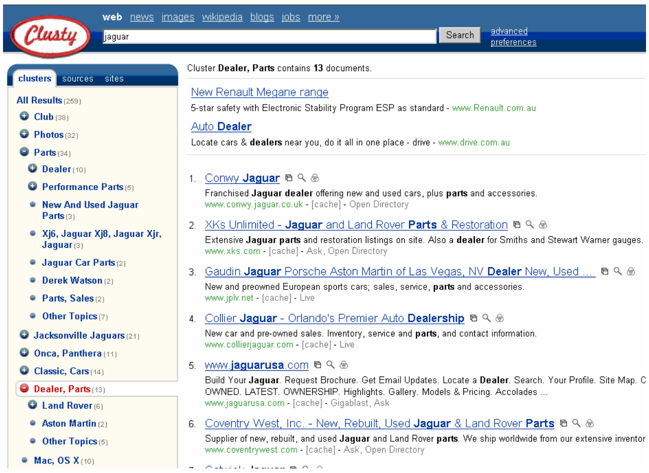
A flat list of search results of most popular search engines delivers no information about knowledge structure related to the returned search results; searched items are isolated from each other and presented to the user independently. For example, when an information seeker is curious about "Self-Organizing Map"[2], with only the flat list of Web snippets, it is very hard for the searcher to grasp what the Self-Organizing Map is for, and which discipline the self-organizing map belongs to. Providing information seekers with an overall view of the hierarchical knowledge structure, and indicating where the returned Web snippets are located in this knowledge structure, is helpful to those who are new to a knowledge domain, and reinforcing to those who are conversant with it.

A flat list format of search results is appropriate when the returned items are less than say 50 (relevant documents reviewed per session are around ten or fewer [3]). Therefore, the thousands, or even millions of search results returned need to be re-organized to facilitate Web information seekers to locate relevant information efficiently.

E. Mismatching Mental Model of Clustering Engines

One approach to re-organize the long list of search results is to cluster the search results based on the cluster hypothesis that relevant documents tend to be more similar to each other than to non-relevant documents [10]. There are already many researchers who attempt to deal with this problem [10][29], and some commercial search engines that cluster search results have been developed. The main problem of clustering search result is that sometimes search results are not properly clustered. For example, Clusty [4], ranked number four of the top 20 search engines by SquirrelNet [25], is a *cluster engine* which organizes search results into folders that group similar items together. When using this search engine to search for "jaguar", the search results are illustrated in Fig 3.

Search results are clustered and organized in a hierarchical structure and presented in groups with subjects/topics. However, from the human being's point of view, the arrangement of the search results is very confusing. For instance, "Parts" and "Dealer, Parts" are two automatically formed top level categories. Under the category "Parts", there is a subcategory "Parts / Dealer". There are 13 items under category "Dear, Parts", and 10 items under the subcategory "Parts / Dealer". All of the 10 items can be found under the category "Dealer, Parts". Re-do the same search 10 minutes later, the subcategory of Parts is renamed to "Parts / Dealer, Cars" and there are only 6 items under this subcategory. Clustering search results and entitling the groups with the extracted topic/subjects like the above example usually do not match the human edited knowledge structure, such *Yahoo! Web Directory*, or the Open Directory Project (ODP) [26]; the automatically formed hierarchy mismatches human mental model.



All Results (259)
- Club (38)
- Photos (32)
- Parts (34)
  - Dealer (10)
  - Performance Parts (5)
  - New And Used Jaguar Parts (3)
  - Xj6, Jaguar Xj8, Jaguar Xjr, Jaguar (3)
  - Jaguar Car Parts (2)
  - Derek Watson (2)
  - Parts, Sales (2)
  - Other Topics (7)
- Jacksonville Jaguars (21)
- Onca, Panthera (11)
- Classic, Cars (14)
- Dealer, Parts (13)
  - Land Rover (6)
  - Aston Martin (2)
  - Other Topics (5)
- Mac, OS X (10)

Cluster **Dealer, Parts** contains **13** documents.

New Renault Megane range
5-star safety with Electronic Stability Program ESP as standard - www.Renault.com.au

Auto **Dealer**
Locate cars & **dealers** near you, do it all in one place - drive - www.drive.com.au

1. Conwy **Jaguar**
   Franchised **Jaguar dealer** offering new and used cars, plus **parts** and accessories.
   www.conwy.jaguar.co.uk - [cache] - Open Directory
2. XKs Unlimited - **Jaguar** and Land Rover **Parts** & Restoration
   Extensive **Jaguar parts** and restoration listings on site. Also a **dealer** for Smiths and Stewart Warner gauges.
   www.xks.com - [cache] - Ask, Open Directory
3. Gaudin **Jaguar** Porsche Aston Martin of Las Vegas, NV **Dealer** New, Used ...
   New and preowned European sports cars; sales, service, **parts** and accessories.
   www.jplv.net - [cache] - Live
4. Collier **Jaguar** - Orlando's Premier Auto **Dealership**
   New car and pre-owned sales. Inventory, service and **parts**, and contact information.
   www.collierjaguar.com - [cache] - Live
5. www.**jaguarusa**.com
   Build Your **Jaguar**. Request Brochure. Get Email Updates. Locate a **Dealer**. Search. Your Profile. Site Map. C OWNED. LATEST. OWNERSHIP. Highlights. Gallery. Models & Pricing. Accolades ...
   www.jaguarusa.com - [cache] - Gigablast, Ask
6. Coventry West, Inc. - New, Rebuilt, Used **Jaguar** & Land Rover **Parts**
   Supplier of new, rebuilt, and used **Jaguar** and Land Rover **parts**. We ship worldwide from our extensive inventor
   www.coventrywest.com - [cache] - Ask, Open Directory

Fig. 3 Search results of "jaguar" from Clusty.com (July 10, 2006)

F. Personalized Search Preferences are Neglected

Personalization, according to [20], addresses the adaptability of Web based information services to the needs and interests of individual users, and thus helps to fulfil one of the main goals of Web sites – the creation of loyal visitors. However, most search engines and tools today try to return and rank search results suitable for general purpose search; personalized search is seldom considered [7]. No matter what role a searcher has - a car salesperson, an environmentalist, or an aircraft enthusiast - if they use the same query "jaguar", they will get exactly the same search results. However, when submitting "jaguar" to a search engine, the salesperson may be concerned only about the jaguar car; the environmentalist is seeking information about animal jaguar; and the aircraft enthusiast is searching for information relating to the jaguar military aircraft. The general purpose search tools do not consider the personal information needs and thus from the perspective of Web searcher's, the *precision* of search results can be very low.

[21] points out that the meanings and resources of Web

[2] A self-organizing map is a type of artificial neural network that is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological properties of the input space. http://en.wikipedia.org/wiki/Self-organizing_map, Nov 9, 2007

search results are usually valued and determined by a group of authors. These authoring biased results are then presented to the entire user population. Web search engines analyze letters and words that make up the contents of documents, and integrate intrinsic document properties such as citations and hyperlinks to the incorporation of usage data. Search results usually contain a specific set of words or meanings by utilizing content-based approaches like statistical or natural language processing techniques. These techniques cannot differentiate which documents are really relevant to users information needs [21].

Using personalized information, on the other hand, can introduce another problem, that is, when a user changes his or her interests frequently. For example, if an environmentalist one day wants to seek some information about a jaguar car, if the search-term is still "jaguar", the personalized search results will mainly be about animal jaguar but not the jaguar car. The situation may become very complex if subsequently the environmentalist is really buying a jaguar car, and further, joins a jaguar car club, becomes a jaguar car fanatic, and then changes his interest to the jaguar aircraft.

G. Low Recall of Web Directory Navigation

Another approach to obtain Web information is browsing or navigating a Web directory, such as *Yahoo! Web Directory* [28] or the ODP [26], by following its hierarchical structure of categories. These Web directories will, like a map, instruct an information searcher where to go to find the relevant information. For example, *Yahoo! Web Directory* has 14 first level categories: Arts & Humanities; Business & Economy; Computers & Internet; Education; Entertainment; Government; Health; News & Media; Recreation & Sports; Reference; Regional; Science; Social Science; and Society & Culture (July 10, 2006). If one wants to find some information about a soccer player, a clear route is: Recreation → Sports → (Type of Sports) Soccer → Players → Men, where a list of male soccer players is presented. If one is interested in *Zidane*, the famous French soccer player, it is easy to find there are six Web sites that are all about the player. By this approach, one can find a place where nearly all the Web pages are relevant; the *precision* of the retrieved results is very high.

However, one serious problem of this approach is the extremely low *recall* of the retrieved results – only six results are listed by *Yahoo! Web Directory* (retrieved on July 10, 2006). If the term "Zidane" is submitted as a query to Google, there are 21,400,000 search results returned, and Yahoo! returns 10,800,000 search results (retrieved on July 10, 2006). Compared with the six Websites given by *Yahoo! Web Directory*, one conclusion easily drawn is that the *recall* of search results of *Yahoo! Web Directory* is much lower than that of Google or Yahoo!'s Web searching. This example also reveals one reason why more and more people are using searching instead of navigating/browsing to retrieve information from the Web. *Yahoo! Web Directory* is maintained by a small group of experts, no matter how hard they work, their edit speed cannot keep up with the increasing growth of the Web. *Yahoo! Web Directory* used to be an essential feature of Yahoo!; however, it is getting farther and farther from the Web information seeker's interest centre.

G. Existing Search Tools are not Integrated

In addition to the general purpose Web search engines such as Google, Yahoo!, and MSN, there are also specific search tool functions which are desirable, for example, desktop search, music search, language specific search, and specific full text database and bibliographic searching. While the specific search tools provide more effective search for a specific domain or field as compared with the general purpose search engines, information seekers must install these tools on their computers, and then match the search tool/function with the information retrieval need. This process may involve considerable trial and error and investment in learning the specific methods of a variety of systems. For example, when searching for a full text academic paper the researcher may first try Google. If Google does not provide full text of the paper, the searcher may try a specific full text database such as the ACM digital library, or a specific e-journal. Jumping from one search engine to another search tool and again to another tool is time consuming, disorientating, and can be discouraging [32].

III. Categorizing/Clustering Search Results to Boost Search Quality

To boost the quality of information retrieval, a search algorithm which combines categorization, clustering and personalization techniques to re-rank and filter the retrieved search results is proposed, as illustrated in Fig 4.

The meta-search engine in the search framework takes as input the information from different information sources, such as the Internet, local networks, commercial databases and personal computers. Search results are first categorized and clustered. Domain ontology in a digital ecosystem is served as a pre-defined knowledge structure for search results categorization, and as a reference ontology for user profile to describe user search preferences. Clustered results are utilized to boost the performance of categorization

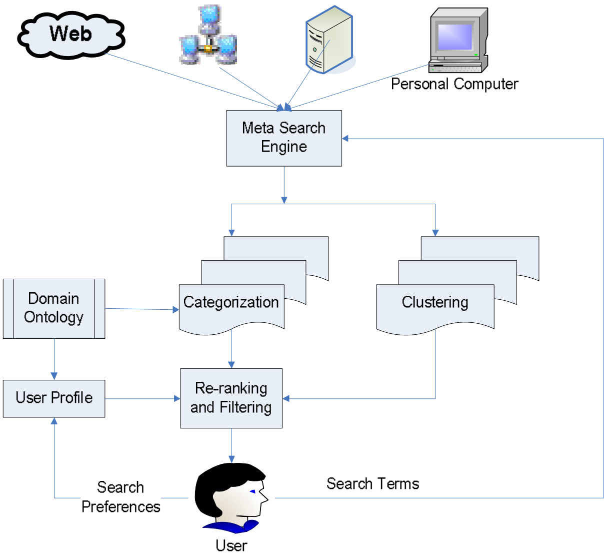


Fig. 4 a search framework in digital ecosystems

[31]. The search results are finally re-ranked and filtered according to learned user profile.

IV. Future Work

The proposed framework is being developed; different categorization and clustering algorithms will be evaluated with regarding to effectiveness and efficiency in this scenario. Tradeoff between effectiveness and efficiency is an important factor to be considered because the categorization and clustering processes have to be implemented dynamically.

V. Conclusion

In this paper, problems of information retrieval in digital ecosystems are presented from a user's perspective, and some of the possible reasons are analyzed and discussed. A search framework is consequently proposed. The approach combines the power of text categorization, clustering, ontology and personalization techniques and we intend to boost the quality of search results in digital ecosystems.

VI. Acknowledgement

This work is partially supported by Digital Ecosystems and Business Intelligence (DEBI) Institute of Curtin University. Thanks to Professor Elizabeth Chang for her continuous support and encouragement.

VII. References

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke and S. Raghavan, "Searching the Web", *ACM Transactions on Internet Technology*, 2001, vol. 1, no. 1, pp. 2-43.

[2] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, New York & Addison Wesley, Harlow, 1999.

[3] G. G. Chowdhury, 2004, *Introduction to Modern Information Retrieval*, 2nd edt, Facet Publishing, London.

[4] Clusty.com, http://www.clusty.com

[5] S. Dumais and H. Chen, "Hierarchical Classification of Web Content", in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 256-263.

[6] P. Ferragina, and A. Gulli, "A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering", in *Proceedings of the Special interest tracks and posters of the 14th international conference on World Wide Web*, 2005, pp. 801-810.

[7] S. Gauch, J. Chaffee and A. Pretschner, "Ontology-based personalized search and browsing". *Web intelligence and Agent System*, 2003, vol. 1, no. 3-4, 219-234.

[8] F. Geraci, M. Pellegrini, P. Pisati and F. Sebastiani, "A Scalable Algorithm for High-Quality Clustering of Web Snippets", in *Proceedings of the 21st Annual ACM Symposium on Applied Computing*, 2006, pp. 1058-1062.

[9] Google Desktop SDK, http://desktop.google.com/dev/indexapi.html

[10] M.A. Hearst and J.O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results", in *Proceedings of the 19th annual international ACM/SIGIR conference on Research and development in information retrieval*, 1996, pp. 76-84.

[11] K. S. Jones, "Document Retrieval: Shallow Data, Deeper Theories, Historical Reflections, Potential Directions", in *Proceedings of the 25th European Conference on Information Retrieval (ECIR'03)*, pp. 1-11.

[12] B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs", *Information Processing and Management*, 2006, vol. 42, pp. 248-263.

[13] B. J. Jansen, A. Spink, and J. Pederson, "A Temporal Comparison of AltaVista Web Searching", Journal of the American Society for Information Science and Technology, vol. 56, no. 6, pp. 559-570.

[14] M. Kobayashi and K. Takeda, "Information Retrieval on the Web", *ACM Computing Survey*, 2000, vol. 32, no. 2, pp. 144-173.

[15] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment", *Journal of ACM*, vol. 46, no. 5, 1999, pp. 604-632.

[16] D.K. Limbu, A. M. Connor, and S.G. MacDonell, "A Framework for Contextual Information Retrieval from the WWW", in *Proceedings of the 14th International Conference on Adaptive Systems and Software Engineering (IASSE05)*, 2005, pp.185-189.

[17] S. Mizzaro, "Relevance: The Whole (hi)story", *Journal of the American Society for Information Science*, 1997, vol. 48, no. 9, 810-832.

[18] S. Osi´nski, "Improving Quality of Search Results Clustering with Approximate Matrix Factorisations", in *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, 2006, pp. 167-178.

[19] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Library working paper, 1998, SIDL-WP-1999-0120 of 11/11/1999.

[20] D. Pierrakos, G. Paliouras, C. Papatheodorou, C.D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: a Survey", *User Modeling and User – Adapted Interaction*, Nov. 2003, vol. 13, no. 4, 311-372.

[21] J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar and T. Breuel, "Personalized Search: A contextual computing approach may prove a breakthrough in personalized search efficiency", *Communications of the ACM,* September 2002, vol. 45, no. 9, 50-55.

[22] G. Ramakrishnanan and P. Bhattacharyya, "Text Representation with WordNet Synsets using Soft Sense Disambiguation", in *Proceedings of the eighth International Conference on Application of Natural Language to Information Systems (NLDB 2003)*, Springer-Verlag, Berlin Heidelberg, pp. 214-227.

[23] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval", *Information Processing & Management*, vol. 24, no. 5, 1988, pp. 513-523.

[24] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, 2002, vol. 34, no. 1, 1-47.

[25] SquirrelNet, http://www.SquirrelNet.com

[26] The Open Directory Project, http://www.dmoz.org

[27] G. I. Webb, M. J. Pazzani and D. Billsus, "Machine Learning for User Modeling", User Modeling and User-Adapted Interaction, 2001, vol. 11, no. 1-2, pp. 19-29.

[28] Yahoo Web Directory, http://dir.yahoo.com

[29] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration", in *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval (SIGIR-98)*, ACM Press, New York, NY, 1998, 45-54.

[30] D. Zhu, "Improving the Relevance of Search Results via Search-term Disambiguation and Ontological Filtering", 2007, Master Thesis, Curtin University.

[31] D. Zhu, "RIB: A Personalized Ontology-based Categorization/Clustering Approach to Improve the Relevance of Web Search Results", Curtin Business School Doctorial Colloquium, 2007.

[32] D. Zhu and H. Dreher, "An Integrating Text Retrieval Framework for Digital Ecosystems Paradigm", in *Proceedings of the Inaugural IEEE Digital Ecosystems and Technologies Conference*, pp. 367-372.