An Integrated Information Retrieval Framework for Managing the Digital Web Ecosystem

Dengya Zhu
School of Management, Information Systems
CBS, Curtin University, Perth, WA, Australia
Email: d.zhu@curtin.edu.au

Shastri L Nimmagadda
School of Management, Information Systems
CBS, Curtin University, Perth, WA, Australia
Email: shastri.nimmagadda@curtin.edu.au

Torsten Reiners
School of Management, CBS
Curtin University, Perth, WA, Australia
Email: T.Reiners@cbs.curtin.edu.au

Abstract

The information explosion constrains the Digital Web Ecosystem exploration and makes challenging in retrieving relevant information and knowledge using Web search tools. The existing tools are not well integrated, and search results are inadequately managed. In this article, we describe effective information retrieval services for users and agents in various digital Web ecosystem scenarios. A novel Integrated Information Retrieval Framework (IIRF) is proposed, which employs the Web search technologies and traditional database searching techniques to provide comprehensive, dynamic, personalized, and organization-oriented information retrieval services, ranging from the Internet, intranet, to personal desktop. Experiments are carried out demonstrating the improvements in the search process with an average precision of Web search results to standard 11 recall level, attaining improvement from 41.7% of a comparable system to 65.2% of search. A 23.5% precision improvement is achieved with the framework. The comparison made among search engines presents similar development with satisfactory search results.

Keywords: digital ecosystem, information retrieval, information management, search engine, crawler

1 Introduction

The digital ecosystem (Gartner 2017) is an interdependent group of elements or processes, such as entities and or dimensions, more specifically people, enterprises, who share standardized digital content on various platforms that are mutually beneficial. Accenture (Moore et al. 2018) finds in a survey with more than 3000 executives that "digital ecosystems are transforming the way their organizations deliver value". However, data in digital ecosystems are distributive, complex, heterogeneous, and multidimensional (Barrows and Traverso 2006). In other words, they have all the features that big data do, that is, the volume of the data generated per hour in digital ecosystems ranges from megabytes to gigabytes to terabytes; tens of thousands of bytes of data transported per seconds, demonstrating velocity of data. The data format varies as per varieties, from emails to instant messages to images to streaming data and more.

Because of the complexity in digital ecosystems, and their data sources, management information systems need improvements, creating scopes for better information retrieval methods and search functions, adaptable to a new information-access era. Searches need highly specialized search tools and formulations in addition to the presentation of search results in a way they can better be interpreted and analysed. Multiple information scenarios of digital ecosystems make businesses shifting their focus to new flexible, re-configurable, and collaborative business paradigm, and consequently, they should adapt to digitalization trends, and strive to leverage their data in competitive advantage. With the vast amount of data in hand, one of the distinct steps for organizations is to facilitate the data search process by easing the complexity of digital ecosystems. For example, Commonwealth Bank of Australia (CBA) developed an app that assists users to search properties that they are interested in with sale price history as well as similar information of other properties nearby, and then with mortgage link points to CBA. After six months, consumers made more than 1.2 million property searches, and the app's return on investment is 109% (Weill and Woerner 2015).

Many enterprise-level search tools/platforms exist for exploring digital ecosystems (Barrows and Traverso 2006). *Search engines*, such as the Apache Solar and Lucene (McCandless et al. 2010) are typical examples of types, which provide services at an enterprise level, such as distributed indexing/searching with high scalability, availability, and extensibility. The customized version may include an entity extractor, thesaurus, classification, filtering and other characteristics. *Desktop search* that focuses on retrieving local files resides in personal desktop computers, messages, emails, Web history, is another type of searching tool. The third type of such gadget is an *intranet search* engine which crawls information from servers within an intranet to local networks. For resolving enterprise-scale problems, a search engine should support and combine tagging, categorization, and navigation tools to improve the end user experience. An enterprise metadata category – ontology-based metadata – can be built to define a metadata schema, to index a set of documents, and write a user interface for querying and displaying results. Even though automatic metadata extraction is never perfect, a user interface is needed to allow amendments and re-use of the metadata. In addition, the system should satisfy scalability, security, metadata update, view privilege, and query (search-term) optimization criteria (Albro 2006). The challenges are identified from the existing literature (Croft et al. 2015) and the search tools.

2 The Existing Issues and Challenges

The information retrieval is challenging, and there are still some big issues need to be addressed (Croft et al. 2015), because of information explosion, low accuracy, search results are improperly managed, mismatching human-computer interaction with clustered results. An active information retrieval service is needed for users in various digital ecosystems. In digital ecosystems, other challenges include inadequately integrated domains, systems and their associated data sources. The current challenges of documentation of data sources and information search are briefly summarized in the following sections.

Integrated search tools: In addition to the general purpose Web search engines, such as Google, Bing and Yahoo, specific search tools and functions exist, for example, desktop, music, language-specific searches, and specific full-text database with bibliographic searching. The specific tools provide a more effective search for a particular domain or field, as compared with the general purpose search engines. Information seekers must install the tools on their computers, and then match the search tools/functions with their information retrieval needs. The process may involve considerable trial and error and investment in learning a variety of systems. Substantial resources are used such as time,

memory, disk space, and processing power, in particular, accessing the high-resolution images. An integrated search framework that can collaborate search function and repository access tools can facilitate resolving the issues, easing information retrieval from search engines.

The syntactic-based search not necessary semantic centric: The searching is syntactical but not semantic based, and search results are not properly personalized. Web search engines look for factual similarities between search-terms and the web-pages (Arasu et al. 2001). Search engines crawl websites from the Internet and download web-pages from various sites. Idioms or phrases are accordingly indexed, stemmed, removing the stop-words in the web-pages. High-level dimensionality document is created as per the content of the web-pages. A number of indexed terms determines the dimensionality. This vector represents different patterns of search-terms. Accordingly, similarities between search-terms and their knowledge including document vectors are calculated and ranked. During this process, the semantic characteristics and issues of search-terms are not fairly involved. As an example, it is not a surprise when using "UPS" as a search term to retrieve information about the "Uninterruptible Power Supply" that may return irrelevant or ambiguous results such as the "United Parcel Services".

Untailored search results: Most search engines and tools make an effort to return and rank search results, based on general purpose search, where a contextual and personalized search is still not widely considered relevant (Croft et al. 2015). No matter what role a searcher is - a car sales representative, an environmentalist, or a computer technician - if they use the same query "jaguar", they get the same search results. However, submitting the search-term, the sales representative thinks of the "jaguar" car and not anything else. The environmentalist seeks information about the animal jaguar, whereas the technician thinks of using the Apple's Jaguar as an operating system. The general purpose search tools thus need improvement to get quality and relevant search results. Although some search engines provide personalized results, the precision still a concern (Croft et al. 2015).

Enterprise-level search personalization: The customization and personalization features go hand-in-hand. Personalization must deliver the content and functionality that match the specific users' needs, what they look for in search of innovative and new terminologies. Netflix is an example, which has established a market with the adaptability of user views and search-terms. Without any clue of search terms, a majority of search engines return and rank search results, based on general purpose search after a user submit a keyword, where personalized search is not considered (Arnold 2004). The keywords may be the characteristic of human language, although the keyword queries are inherently ambiguous. For example, keyword "jaguar" may mean 1) a car for a car seller; 2) an animal for an environmentalist and 3) an operating system for a technician. The personalized Web search tools and models are taken advantage of, even though development of the search tool is still in its initial stages that may need the attention of real-world applications.

Integrated holistic enterprise-level search engine: Enterprise software applications typically hold their search functions. More popular Gmail, Microsoft Outlook and other email services have built-in search functions. Microsoft SharePoint built-in search allows users to search SharePoint pages. Nevertheless, users in an enterprise need to frequently change several portals and explore different parts of data available in the enterprise.

The research gaps are identified through critical analysis of the existing facts of search tools (Zhu and Dreher 2007), diligently providing evidence and substantiating the results. Further, we examine them by exploring and framing research questions and objectives with problem solutions.

3 Research Questions and Objectives

Based on the current information retrieval challenges, we design the research question and objective. Research questions (RQ): (1) Design and develop a framework in search of structured and unified information (2) How do we accomplish information search in complex digital Web ecosystems and analyse their results, using the latest search engines and tools. The research objectives (RO): (1) Develop an integrated information retrieval framework that can deal with the search engines and (2) Analyse the search results to claim that the research performs well with user satisfaction.

4 Research Goal, Motivation and Significance

The goal is to design and develop a framework to search for unified information and validating the framework through series of experiments with supported performances. Another research goal is to generate structured textual data that can be shared between different search engines and the type of framework needed for such unified information. Increased search efforts and outcomes of terms searched in various contexts is the motivating factor for exploring and exploiting new framework. Keeping in view the user needs, variety of search engines and options, a need for more generic framework is felt. Activation, persistence and intensity of search terms and their outcomes usable by type of users motivate us to develop new search articulations. Connecting and trending of events between contexts are the other motivations. The implication, relevance and quality of articulations in the current contextual applications, all describe the significance of research. The current research is made significant to academicians, IT/IS researchers and data management personnel. Integrated methodological framework can risk minimize the economics involved in exploring new information and knowledge through interfaces and desktop integration. Based on the experiences with IT/IS companies, the authors experience pitfalls at various stages and chains of search engines, especially keeping in view the heterogeneity and multidimensionality of data sources relevant to industry scenarios.

5 An Integrating Search Framework

Figure 1 illustrates an Integrating Information Retrieval Framework (IIRF), proposed for digital ecosystem representation, articulated with various artefacts. It is detailed in the following sections.

5.1 Information streams

When users submit search requests, search-terms match with information from diverse sources including the Internet, intranets, full-text databases, databases of digital ecosystems, and personal desktop computers. The similarities are compared between search results, and the categories are sorted through adaptable ontologies associated with digital ecosystems. The similar and dissimilar items, obtained from different sources are clustered. The search processes are described in the following sections, with the functionality of each component, as explained in Figure 1. The framework is divided into four main parts, namely, information streams and sources; query or search-term expansion; search results categorizing/clustering and filtering; and personalized search results representation. The components are described in the following sections.

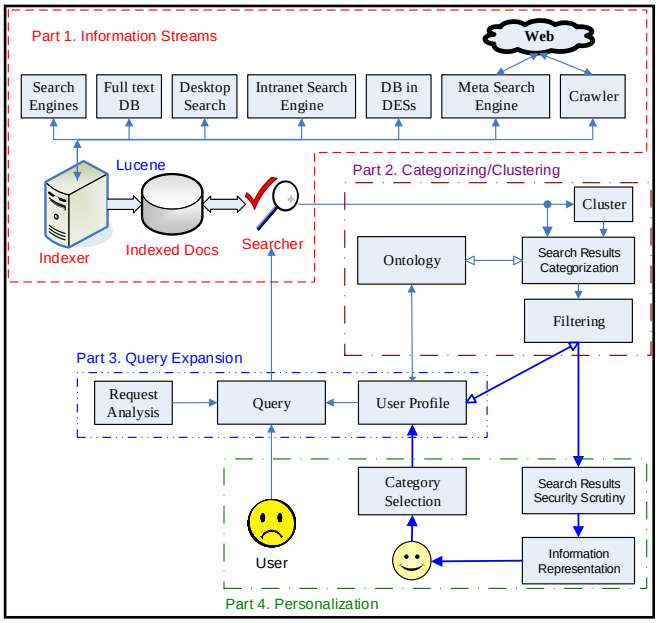


Figure 1: An integrating information retrieval framework (IIRF)

Search process: A programmable script is designed as a crawler for browsing the World Wide Web. Currently, the crawlers emerge with large-size scripts, complexity and rapid growth of the World Wide Web. With the result, the page selection, importance, recency, and refresh options create many challenges. In addition to regular page refresh, crawlers should focus on the effectiveness of crawling. Usually, a crawler starts with an initial set of URLs that are placed in a queue and prioritized. A URL is selected based on some ordering strategies. The crawlers download Web-pages, extract URLs from the downloaded pages, and put the new URLs in the queue, expanding into (crawling) relevant Websites.

This process is repeated until crawlers decide to abort. Prioritizing the URLs in the queue and setting the stop conditions are both related to estimating, or measuring the relevance of the URL content to the semantic need of digital web ecosystems.

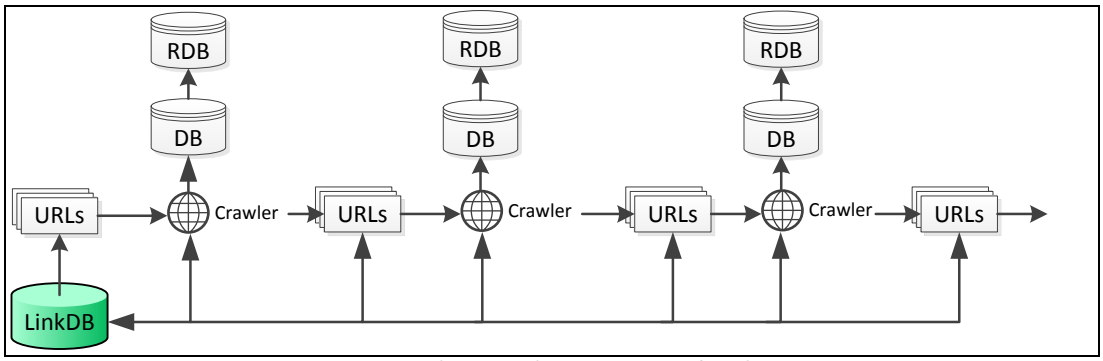


Figure 2: The crawling strategy for the IIRF

As illustrated in Figure 2, the proposed crawling strategy consists of the following stages. Firstly, a set of seed URLs, which contains all relevant websites picked up by ecosystem users, is created and stored in a URL link-database (LinkDB), injecting in the URL list (URLs, also URL Frontier). The seed URL list is adjustable based on users' business requirements. With the seed URLs injected into the URL frontier, a crawler downloads all Webpages in the URLs and stores the extracted content into a database (DB). The content gets indexed by search engine after the documents are preprocessed (stop word removing, stemming, named entity extraction). At the same time, URLs contained in the downloaded webpages are also extracted and then put into LinkDB after duplicate and irrelevant URLs are removed. For all the downloaded webpages in the DB, some of them may not be relevant. Therefore, a document categorization algorithm such as Support Vector Machines, is trained by using existing documents stored in the ecosystem as a training data set. Only the relevant documents moved from DB to relevant database (RDB) are indexed in the RDB.

Search aggregators: The users and their adoptable search engines are valued as long as they satisfy the user needs and revenue generated. A generic architecture is needed to match multiple search options and opportunities for Meta integration purposes. A metasearch engine is a system that provides unified access to several existing search engines (Meng et al. 2000). The aggregator in IIRF is based on the following considerations: 1) single search engine's processing power may not scale to the tremendous increase and virtually unlimited amount of data; 2) it is hard or even impossible for a single search engine to index all the data on the Web and keep it up to date; and 3) some 'deep web' sites may not allow their documents to be crawled by external websites, but allow their documents to be accessed by their search engines (Zhu and Dreher 2007).

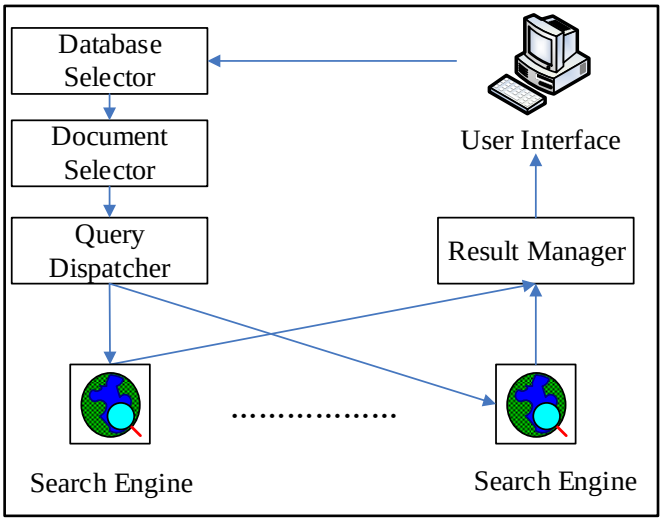


Figure 3: Meta-search engine structure (Meng et al. 2000)

The conceptual architecture of the meta-search engine is illustrated in Figure 3. Meta-search starts with

initializing user query, selecting a set of suitable databases (coupled with search engines) by the database selector. The document selector chooses the number of documents, retrievable from the component search engine. The local similarity threshold is used to limit the documents retrieved from the component search engine. Query dispatcher connects the server with each of the selected search engine, and passes the query to them. Results manager integrates returned results from search engines into a single ranked list and renders it to the user (Meng et al. 2000).

Reprising databases: A Relational Data Base Management System (RDBMS) usually manages the existing databases in an enterprise that can be repurposed by IIRF in the digital Web ecosystems. In addition, the RDBMS manages the metadata and other security-sensitive data, which usually provides security and integrity management (Ramez and Navathe 2016). In IIRF, user queries are submitted to RDBMS as well, and the retrieved results from the RDBMS are then presented to the categorization/clustering component for further processing, as illustrated in Figure 1. In addition, the extracted entities such as title, gender, person name, address, email address, organisation, telephone number, mobile phone number and others from text document are stored in the RDMBS. These extracted entities are used to match records of RDBMS, and thus a connection between structured data in RDBMS and unstructured/semi-structured text data can be established, and verse vice. Users of IIRF can search data by providing a Google like search term, and can also search data in the RDBMS.

Desktop search: Desktop integration feature must be agreeable among users to share data between applications, sustainable to business leverages. The Google, Microsoft, Yahoo!, and other significant players (Barrows and Traverso 2006) provide free downloaded desktop search solutions. As the storage allows hundreds of gigabytes of data and beyond, the desktop search can significantly make better user productivity. In these models, indexing or tagging can enable users to access information through dynamic integrated navigational retrieval systems. They can act as agents to return pointers or links to the desired information. The plain list representation, and no consideration of semantic aspects, though are drawbacks but overcome by re-organising, ontologically filtering and personalizing the desktop search results.

Text retrieval: Text retrieval differs from database probing, in which an exact matching is necessary, and any mismatching among thousands of objects can create an error. However, there is more tolerance for errors, because of information retrieval systems focus on process semi- or unstructured natural language text. The text in such contexts could exhibit ambiguities semantically (Croft et al. 2015). The added difference is that in a full-text retrieval system, the search is about subject or topic information, which involves not only the syntactic interpretation of search-terms and text objects but also the relevance of an object to user information needs. Some open source search engines, such as Lucene are available for full-text search.

Intranet engine: Searching for information on the intranet is rather a daunting task, which is addressed by search tools' development companies, such as Google and Thunderstone. Because the performance of intranet search engines differs dramatically, ISYS Search Software vendors suggest that features should be considered carefully, and in IIRF contexts, intranet engine selection is the users' choice.

5.2 Search results categorization and clustering

Knowledge-based hierarchical ontologies are considered to manage a vast amount of information. Another method is to arrange the itemized information into different clusters according to their similarities. In this retrieval framework, these two approaches are combined to leverage the advantages of both tactics, as suggested in Chau and Chen (2008).

Fine-grained ontologies: The ontologies are created in IIRF to improve the search process, categorizing the search results. The search results are filtered based on the user selection and accordingly classified under the selected category for description and presentation. In IIRF, the Open Directory Project (ODP) is employed as an ontology to present the Web knowledge structure. The semantic characteristics of each category in the ODP are manifested by a category-document that includes the topic of the class, the description of the type, and a list of submitted Web-pages (composed of the title of the Web-pages and a brief description of each of the submitted Web-pages) under this category.

Categorizing search results: Undoubtedly, the text categorization is the problem of assigning predefined categories to free text documents (Croft et al. 2015). In IIRF, search results are categorized

based on the ODP as a lightweight ontology. The category-documents in the ODP are employed as a training data set.

Let $d_j = \{w_{1,j},\ w_{2,j},\ w_{T,j}\}$ is the jth category-document, where T is the total number of vocabulary in IIRF, and $q = \{w_{1,q},\ w_{2,q},\ w_{T,q}\}$ is a searched item. $w_{i,j}$ is the tf-idf weight of ith term in jth document, $w_{i,q}$ is the tf-idf weight of ith term in the searched item. The similarity between q and d_j is estimated by the cosine value of the angle ɵ of the two vectors:

$$sim(d_j, q) = \cos(\theta) = \frac{d_j \bullet q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{T}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{T} w_{i,j}^2 \sum_{i=1}^{T} w_{i,q}^2}}$$

The similarities between q and the d_j, $j = 1, 2 \ldots N$ (where N is total number of *category-document* in IIRF) are ranked/sorted in their descending order. For top K ranked *category-documents*, suppose their corresponding ODP category is $C = \{c_1, c_2 \ldots c_K\}$, q is assigned to the category selected from C by majority voting algorithm.

Clustering: Text clustering aims at assembling documents that are related among themselves and satisfy a set of characteristic properties. It can be used to expand a user query with new and related index terms (Croft et al. 2015) and facilitate users to browse the retrieved results. Many clustering algorithms are available, such as the K-mean clustering algorithm, and Fuzzy C-Means. In IIRF, K-mean is chosen to cluster returned search results. Two essential issues of K-means are, how to decide a proper K and how to select the original K cluster centres. In IIRF, since the search results are categorized based on the OPD category, the number of the first level categories under which search results are assigned is a suitable candidate for K. Meanwhile, the search-item which is most similar to the candidate category is assigned to the first centre of the cluster. Cosine similarity (as described above) is utilized to estimate the "likelihood" among neighbouring groups.

Filtering: The Google and Facebook introduce personalization features and algorithms that filter information as per user requirements and influence the filtering process. We further analyze filtering processes to show how the personalization can be linked to filtering techniques without any bias of human and computer interaction. Search result filtering is decided in IIRF by two factors: one is the user's selection of an exciting category of the ontology; another factor is the pre-built user profile that is to be discussed in next section. When a user chooses an interesting category, only search results categorized under this category are presented to the user, and other information is filtered out. However, by default, even if the user does not select a type, the search results are filtered. Based on the pre-built user profile, search results are compared with the features in the pattern; they are re-ranked and then only search results having similar features described by user profiles, are presented to the users.

5.3 Expanding and analyzing queries

User profile and personalization: A user profile is a reference ontology in which each concept has a weight indicating the user's interest in that concept (Croft et al. 2015). An information space of the ODP (Pitkow et al. 2002) is used to represent user models. As suggested in Dolog and Nejdl (2003), the user model combines two proposed standard learner profiles, IEEE Personal and Private Information and IMS Learner Information Package (LIP) to express the features of a user. The precision and recall of metasearch engines are improved through mining association rules that reflect users' past search behaviour. IIRF cuts off the ODP knowledge from the second level to obtain 573 topics and uses these topics to represent users search interests that are represented by <topic, weight> tuples. The user profile is initialized by asking users to assign a weight (integer) to an existing topic to indicate how much interesting the topic is. Users are allowed to choose any number of interesting topics. To map a user search interest into these topics, for each search result r_i visited by a user, let c_i is the topic, as detailed in Section 2, the corresponding weight of c_i in the tuple is increased.

Query and request analysis: Query augmentation and result processing are two primary uses of user profiles. In IIRF, after a user selects an OPD topic, query augmentation is an alternative, which allows

the user to re-submit the query q+ = q ∪ {t_k, k = 1, 2 …K}, where q is the query submitted by the user, t_k is the term selected using

$$\lambda^2 = \frac{N[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_k)P(\bar{c}_k)}$$

Where N is the 573 topics in user profile. These topics are now represented by *category-document*. P(t, c) is the joint probability of term t and category c, $\bar{c}$, $\bar{t}$ indicates c or t does not appear respectively. K is number of terms determined by the confidence P(λ^2 > 10.83) < 0.001 = 99.9% that the assumption of independence of the term t_k and q can be rejected (Manning et al. 2008).

5.4 Representing personalized search results

Information representation: In IIRF, search results are obtained from the traditional database, full-text database, intranet, Internet, and desktop searches - all results are integrated into one coherent information representation. Users of the basic search framework can choose which data sources are to be used, from which the needed information is retrieved. The IIRF permits users to set the search scope and thus provide the flexibility to access data sources to satisfy their information needs. Search results come from the Web and Intranet are categorized into a domain-oriented knowledge structure (Zhu and Dreher 2007).

Search results from security scrutiny: The component under development, performs a search result. A security scrutiny task concerns "who is allowed to update a piece of metadata and view a particular piece of metadata about a document (or know that the document exists at all)" (Barrows and Traverso 2006).

6 Experimental Results

Evaluation measurements: The two widely accepted measurements of information retrieval effectiveness are precision and recall. Recall measures the ability of an information retrieval system to retrieve all relevant documents; accuracy measures the ability of an information retrieval system to extract only relevant material. For text categorization purposes, the two measures are defined as (Manning et al. 2008):

$$recall = \frac{categories\ \ found\ and\ correct}{total\ \ categories\ correct}$$

$$precision = \frac{categories\ \ found\ and\ correct}{total\ \ categories\ \ found}$$

In the search process, search engines make available for retrieved documents in a ranked list according to the degree of relevance of the document to a given query. Users then examine the ranked list starting from the top document. The recall and precision measures vary since the users proceed with their examination of the retrieved answer set. Precision versus recall curve is drawn, evaluating the ranked lists. It is based on eleven standard recall level, as usually employed in (Croft et al. 2015). Another measurement is the precision at a given cut-off level. A cut-off level is a rank that defines the retrieved set. For example, a cut-off level of 10 represents the top ten retrieved documents in the ranked list. If seven out of ten of the returned documents is relevant, the precision at cut-off level ten (P@10) is 7/10 = 0.7 = 70%.

Experiment description: A unique search browser (SSB) (Zhu and Dreher 2007), which is a component of IIRF, has been developed which categorizes Web search results from Yahoo! under the categories of the ODP. Five search-terms with general or ambiguous meaning are selected as shown in Table 1.

For each search-term, 50 search results are retrieved by utilizing the Yahoo! Search Web Service API. The returned search results are presented to judges to perform the relevance judgment. The relevant judgment results are summarized. A final binary decision is made regarding whether a returned search item is related to the specified information need or not. Based on the relevance judgment results, the standard 11 recall-precision curve is drawn for each search-term of the search results of Yahoo!, and of the categorized results of SSB, as shown in Figure 4. P@5 and P@10 are two sets of search results, as presented in Table 2.

Query	Information needs
Clinton	The American president William J Clinton
Ford	Henry Ford, the founder of Ford motor company
Health	State of physical, mental, and social-well being
Jaguar	Information about an entity, animal "jaguar"
UPS	Information about "uninterrupted power supply"

Table 1. Search Terms and Information Needs

Company	P@5	P@10	Average
Yahoo	46.7	42	44.4
SSB	85	70	77.5
Improvement	38.3	28	33.2

Table 2. P@5 and P@10 of Yahoo and SSB

%	Goggle		MS Live		Lycos		Bing	
	P5	P10	P5	P10	P5	P10	P5	P10
Clinton	40	40	0	0	40	40	20	10
Ford	20	30	20	20	20	20	40	20
Health	100	100	60	40	100	80	100	70
Jaguar	20	40	20	10	20	40	20	10
UPS	20	10	20	20	20	10	0	0

Table 3. P@5 and P@10 of Google, MS Live Search, Bing and Lycos

The data documented in Tables 2 and 3 are based on different relevance judgment criteria and the summary of estimation done in the current research. Following the macro-averaging style (Manning et al. 2008), while drawing the standard 11 point recall-precision curve, the precision p_j at recall level *j* is calculated by:

$$p_j = \frac{1}{N}\sum_{i=1}^{N} p_{i,j} \qquad j = 0, 1, ..., 10$$

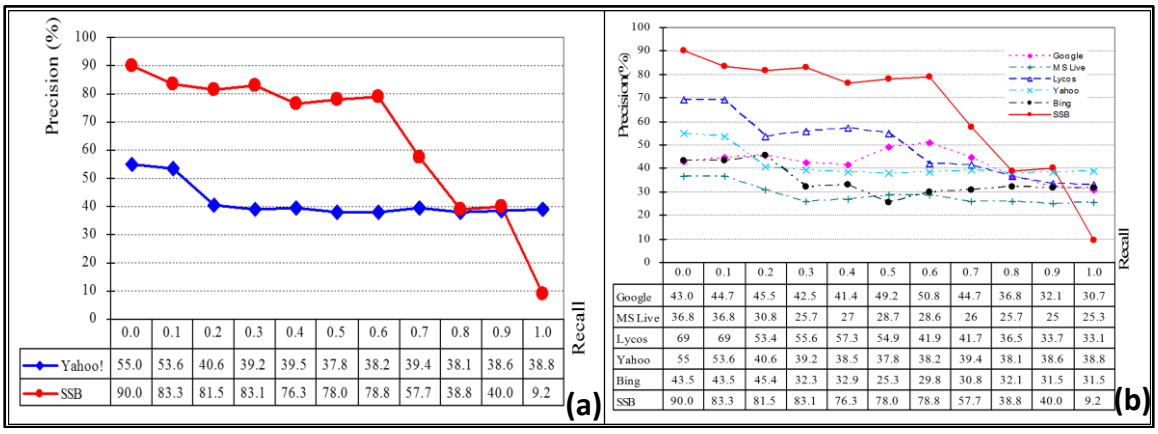
N = 5 is number of queries in the experiment, $p_{i,j}$ is the precision of the *i*th query at *j*th recall level. The overall precision is calculated by:

$$p = \frac{1}{11 \times N}\sum_{j=0}^{10}\sum_{i=1}^{N} p_{i,j}$$

Relevance judgment: Relevance judgment is inherently subjective (Alonso and Mizzaro 2009). To alleviate the subjectiveness introduced, five judges from Curtin University of Technology are involved in this research. They are presented with search-terms, the information needs, and the search results returned from Yahoo! They know nothing about the categorized results of the SSB. The relevance judgment results of five judges are then averaged. A final binary decision is reached for each returned search item of Yahoo!

The experimental results in Figure 4a and Table 2 of SSB are based on 50 search results of Yahoo!, that is, SSB categorized Yahoo's 50 results into different ODP categories. In this context, the comparison between Figure 4a and Table 2 is a direct analogy. Because SSB does not categorize the search results of the rest of four search engines, and in such case, the comparison may not be straightforward. However, the indirect comparison also reveals that without applying the proposed search strategy, the performance of the search engines is far from satisfactory regarding the recall-precision curve measure,

and P@5 and P@10 evaluations.



	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Yahoo!	55.0	53.6	40.6	39.2	39.5	37.8	38.2	39.4	38.1	38.6	38.8
SSB	90.0	83.3	81.5	83.1	76.3	78.0	78.8	57.7	38.8	40.0	9.2

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Google	43.0	44.7	45.5	42.5	41.4	49.2	50.8	44.7	36.8	32.1	30.7
MS Live	36.8	36.8	30.8	25.7	27	28.7	28.6	26	25.7	25	25.3
Lycos	69	69	53.4	55.6	57.3	54.9	41.9	41.7	36.5	33.7	33.1
Yahoo	55	53.6	40.6	39.2	38.5	37.8	38.2	39.4	38.1	38.6	38.8
Bing	43.5	43.5	45.4	32.3	32.9	25.3	29.8	30.8	32.1	31.5	31.5
SSB	90.0	83.3	81.5	83.1	76.3	78.0	78.8	57.7	38.8	40.0	9.2

Figure 4 (a): Average recall-precision curves of Yahoo! and SSB categorized search results over the five search-terms (b) Average recall-precision curves of Yahoo!, Google, MS Live Search, Bing, Lycos, and the categorized results of IIRF

Standard 11 Recall-Precision curve, P@5 and P@10

- As shown in Figure 4b, the overall average *precision* of the 50 search results Yahoo! is 458.9/11 = 41.7%.
- The overall average *precision* of the categorized results is 716.7 / 11 = 65.2%
- The average improvement on *precision* of the Categorized results is 65.2% – 41.7% = 23.5 (%)
- Table 2 demonstrates the improvement of P@5 and P@10 respectively are 38.3% and 28.0% respectively, the average improvement is 33.2%.

Search results of Google, MS Live, Bing, and Lycos: To further verify the effectiveness of IIRF, search results of five queries from Google, MS Live Search, MS Bing, and Lycos are also compared, as shown in Figure 4b. The P@5 and P@10 of the results are shown in Table 3 (where P5, P10 represent P@5, P@10 respectively). Figure 4b and Table 3 reveal 1), SSB outperforms the other five major search engines regarding averaged precision based on standard 11 recall-precision curves and P@5 and P@10 in this experiment. 2), the performance of Google and Yahoo is nearly the same; MS Live Search and MS Bing perform relatively weak, as both the recall-precision curve and P@5 and P@10 demonstrate better search. 3), Lycos performs better by the measure of the recall-precision curve, but only better than MS Live Search and Bing when evaluated by P@5 and P@10.

Although MS Bing classifies its results into different clusters, for the five search terms, only one formed cluster "Bill Clinton" is relevant. For the rest four queries, the formed clusters are irrelevant to the specified information needs. For example, when search results of "jaguar" are clustered, all formed clusters are about cars: Jaguar Cars, Jaguar XF, Jaguar UK, Jaguar Dealers, Jaguar Accessories, BMW, Mercedes, and Jaguar X Type. The framework validates the case studies with variables used in the search results, in addition to suggesting significant rework and validation with more case studies making more reliable search results.

7 Conclusions

In this paper, a novel search framework is proposed, aiming at providing effective information retrieval services in digital Web ecosystems. The search framework integrates not only traditional database search (MIS) and Web search (search engine), but also intranet search, desktop search, full-text database search, personalization, ontological search results categorization, search results clustering, and search results from security scrutiny. Experimental data demonstrate that text categorization based on the ODP can potentially improve the precision of Web search results by 23.5%. Further work is needed to complete the whole framework and conduct evaluation studies, especially to leverage the advantages of information categorization and clustering.

8 Future Scope and Limitations

The experimental results so far demonstrate that text categorization in IIRF improves the precision of Web search results. Implementing the rest of the search framework and conducting a wide range of experiments are planned. With the improvement of precision, the recall of categorized search results has lower than the search results of Yahoo! The reason for the issue is, firstly one search result is assigned to only one category, even if the second or third ranked category has very close similarity instance with the chosen one. Another reason is the categorization algorithm utilized in the IIRF is not optimal. However, algorithms are more powerful in multi-label categorization strategies and their implementations. Combining text clustering and categorization is likely the next research direction, which can improve the recall of the categorized Web-search results.

References

Albro, E. E. 2006. Google Mini Is a Mighty Search Tool," *PC World,* June 21.

Alonso, O. and Mizzaro, S. 2009. Relevance criteria for e-commerce: a crowdsourcing-based experimental analysis, proceedings of the 32nd international ACM SIGIR conference on research & development in information retrieval, p. 760-761.

Arnold, S. E. 2004. How Google Has Changed Enterprise Search," *Searcher,* vol. 12, no. 10, 2004, pp. 8-17.

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

Barrows, R. and Traverso, J. 2006. Search Considered Integral," *ACM Queue,* May 2006, pp. 30-36.

Chau, M. and Chen, H. 2008. "A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis," Decision Support Systems, vol. 44, no. 2, Pages 482-494.

Croft, W.B., Metzler, D. and Strohman, T. 2015. Search Engines – Information Retrieval in Practice, Pearson.

Dolog, P. and Nejdl, W. 2003. Challenges and Benefits of the Semantic Web for User Modelling," in Proceedings of the 12th International *World Wide Web Conference (WWW'03)*, 2003, pp. 99-111.

Gartner, "Insights From the 2017 Gartner CIO Agenda Report: Seize the Digital Ecosystem Opportunity," 2017.

McCandless, M., Hatcher, E. and Gospodnetić, O. 2010. *Lucene in Action*, 2nd, Greenwich: Manning Publications.

Manning, C. D. Raghavan, P. and Schütze, H. 2008. *Introduction to Information Retrieval*, Cambridge: Cambridge University Press, 2008.

Meng, W., Yu, C. and K.-L. Liu, K. –L. 2000. Building Efficient and Effective Metasearch Engines," *ACM Computing Surveys,* vol. 34, no. 1, 2000, pp. 48-89.

Moore, R., Seedat, Y., and Chen, J. Y. J. 2018. South Africa: Winning with Digital Platforms, Accenture, 2018.

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

Ramez, E. and Navathe, SB. 2016. Fundamentals of Database Systems, Pearson.

Weill, P. and Woerner, SL. 2015. Thriving in an Increasingly Ditigtal Ecosystem, MIT Sloan Management Review, Vol. 56, No. 4, pp 27-34.

Zhu, D. and Dreher, H. V. 2007. *An integrating text retrieval framework for digital ecosystems paradigm, Proceedings of the Inaugural IEEE, DEST,* Cairns, Australia, pp. 367-372.

Acknowledgement

We acknowledge the contributions of Professor Heinz Dreher (retired) for his critical comments made on the initial manuscript. We are thankful to the Head of School of Management and Information Systems' Discipline for permitting us to present and publish in the ACIS 2018 conference proceedings.