Journal of Web Engineering, Vol. 12, No.1&2 (2013) 041-064

A SURVEY OF FACETED SEARCH

BIFAN WEI JUN LIU QINGHUA ZHENG

SPKLSTN Lab, Department of Computer Science

Xi'an Jiaotong University

weibifan@sohu.com, [liukeen, qhzheng]@mail.xjtu.edu.cn

WEI ZHANG

Amazon.com, Inc

wzhan@amazon.com

XIAOYU FU BOQIN FENG

Department of Computer Science

Xi'an Jiaotong University

yykfxy@yahoo.com.cn, bqfeng@xjtu.edu.cn

Received November 3, 2011

Revised November 15, 2012

Faceted Search is an exploratory search mechanism, which provides an iterative way to refine search results by a faceted taxonomy. With the benefit of search results diversification, no need for a priori knowledge, and never leading to zero result, it can significantly reduce information overload. Faceted Search has witnessed a booming interest in the last ten years. In this paper, we first analyze the representative facet search models. Next, we present a general faceted search framework, and survey the related methods and techniques, including facet term extraction, hierarchy construction, compound term generation and facet ranking. Then we discuss the metrics for faceted search evaluation, and also highlight the main characteristics of a number of existing faceted search systems. Some directions for future research are finally presented.

Key words: Facet, Faceted Search, Faceted Taxonomy, Metrics

Communicated by: B. White & M. Norrie

1 Introduction

Search engines currently have become the indispensable tools for web users to locate information. Although widely used, most search engines still suffer from the following issues, and fail to meet

certain requests from users. First, the keyword search interface has the "vocabulary problem" [1]. The keywords used in a query might be different from those indexed by the search engines. The "vocabulary problem" leads to the mismatch between the search results and users' needs. Secondly, most search engines use "one-list-only" approach to represent search results. It blends the search results of different topics into a single result list [2, 3], which is often too long and daunting for searchers. Thirdly, conventional search engines adopt a "trial-and-error" approach that lacks the progressive filtering mechanism [4]. All of these result in the information overload problem [5]. That is to say, users need to spend a great deal of efforts to select the information needed from the search results list.

Two types of approaches, search results ranking and diversification, have been applied to address the information overload problem [6]. By using relevance functions, the ranking approach puts the most relevant results to the top of a result list. However it may not perform well when the query is too general and the result set is too large, since it will be difficult to judge which result is better than the other. The diversification approach groups the search results into a wide variety of categories according to keywords, tags or other metadata. Thus, users can only focus on the results of an interested category, ignoring other results, which effectively alleviates the information overload problem. The diversification approach can be further divided into sub-categories, including taxonomy-based method, controlled vocabulary based, and thesauri-based method, faceted classification, and clustering [7]. Among these, the search paradigm using faceted classification is known as faceted search.

Faceted search is an exploratory approach, which provides an iterative way of refining the search results by facets. In recent years, faceted search has been an area of intense investigation[4, 8-10], and is widely applied to e-commerce sites (eBay [11], Amazon [12]), bibliographic databases (Faceted DBLP [13], IEEE/IET Electronic Library [14], ISI Web of Knowledge [15]), multimedia libraries (Open Video Digital Library [16], Google Images [17]) and other fields.

In this section, we first introduce some key concepts of faceted search. Then conduct a comparative study of faceted search and other three search paradigms. Finally, a faceted taxonomy of research work on faceted search is briefly explained.

1.1 Key concepts in faceted search

Figure 1 shows a screenshot of the user interface of ISI Web of Knowledge service [15], which provides an intuitive impression of faceted search. We will use it as an example to discuss the concepts of facet, faceted taxonomy, and faceted search respectively.

a) Facet. This concept was firstly introduced by S.R. Ranganathan [18] to describe the multidimensional properties of documents. Ranganathan proposed five fundamental facets including Personality, Matter, Energy, Space and Time (PMEST), and developed the first faceted classification system Colon Classification.

After that, many literatures have proposed to refine the original definition of facet. For instance, Prietodiaz [19] defined a facet as one dimension or aspect of a subject. Spiteri [20] defined facet as a group of terms about a specific aspect of a subject, and there should be no common term between any two facets. Each term in a facet represents an attribute or a category, which were also referred as

"*attribute*," "*faceted metadata*," "*facet term*," and "*faceted value*" in other literatures [3, 9, 21-23]. For example, in Figure 1, the "*General Categories*" facet consists of a predefined set of terms including "*SCIENCE & TECHNOLOGY*," "*SOCIAL SCIENCES*," "*ARTS & HUMANITIES*," and the term containing multiple words is defined as compound term such as "SOCIAL SCIENCES." The structure of a facet can be either flat or hierarchical [10, 24], which indicates the terms in a facet have no relation, or have hierarchical relations, respectively.

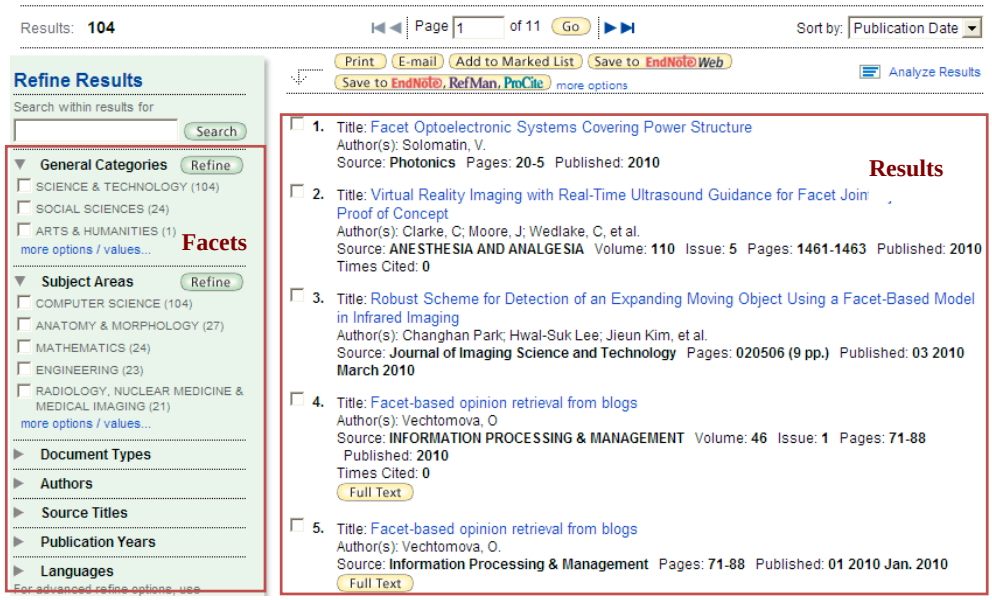

Figure 1. The faceted search interface at ISI Web of Knowledge

b) Faceted taxonomy. A faceted taxonomy comprises a group of taxonomies, each describing one facet [25]. As an example, the faceted taxonomy shown in Figure 1 includes the facets of "*General Categories*," "*Subject Areas*," "*Document Types*," "*Authors*," "*Source Titles*," "*Publication Years*," "*Languages*," and so on. A faceted taxonomy can be used to identify user intents [26] and instant overviews [27].

In a faceted taxonomy the terms of different facets are orthogonal. It means one term cannot appear in multiple facets [28]. This ensures the separability of a faceted taxonomy. That is to say, if the terms or structure of one facet were changed, other facets would not be affected.

c) Faceted search. Faceted search, also referred as faceted navigation or faceted browsing, is an interactive, heuristic and progressive refinement search paradigm. Based on a faceted taxonomy, faceted search allows searchers iteratively select facets and facet terms to narrow down the search results [22, 29-33].

In practice, only a subset of a full faceted taxonomy is presented in the interface. They are dynamically generated based on the results of each iteration of search [34]. The dynamic taxonomy provides not only the summary of current search results, but also a set of links leading to new search results [35].

1.2 Comparison between faceted search and other search paradigms

Other than the faceted search, there are three other widely used search paradigms:

a) Keyword search. It uses an one-search-box interface to obtain user keywords; and the search results are expressed in an one-result-list-only manner. Google, Yahoo, Bing and other mainstream search engines adopt this approach.

b) Form-based search. This approach provides a more advanced query interface to perform complicated searches. By using multiple query fields form-based search is more flexible and easier to use compared to keyword search. Form-based search interface is mainly adopted by hidden websites, such as US National Science Foundation [36] and Hotelbook [37].

c) Directory search. This approach employs a monolithic taxonomy for navigational search [30]. Unlike faceted search, every data item only belongs to one category of the monolithic taxonomy in directory search. Several portal websites, such as Yahoo! Directory [38] and Open Directory [39], have adopted this approach.

We conducted a comparative study on faceted search with keyword search, form-based search and directory search in the round. The main characteristics of these search mechanisms are listed in Table 1.

Table 1 Comparison between faceted search with other search paradigms

Items	Faceted Search	Keyword Search	Form-based Search	Directory Search
Search Interface	**Faceted taxonomy:** 1) Using multidimensional taxonomies for satisfying variant search needs [21] 2) Dynamic taxonomy 3) Using a mouse clicking for navigation [21]	**Keyword:** Suffering from "vocabulary problem"	**Form:** "Field-by-field" search interface, providing multiple query options	**Monolithic taxonomy:** 1) Static taxonomy 2) Using a mouse clicking for navigation 3) Unable to adapt to the thought patterns of different searcher [40]
Prior Knowledge	Requiring little prior knowledge of data schema [41]	Requiring prior knowledge of the dataset to be searched	Similar to faceted search	Similar to faceted search
Navigation Function	1) Refining search results using different facets 2) The number of data items in each category can be used for next navigation [42] 3) Leading to non-empty results [43]	"Trial-and-error" interaction [34], having no navigation function	Similar to keyword search	Monolithic taxonomy may lead to "disorientation" problem
Diversification Function	Diversifying search results using only a small number of facet terms [44]	"One result list only," not supporting search results diversification	Similar to keyword search	Requiring much more terms to achieve similar diversification effect of faceted search
Ranking Function	Supporting facet ranking and search results ranking	Only supporting search results ranking	Similar to keyword search	Similar to keyword search

From the above comparison, it can be concluded that the usability of faceted search in general is better than that of other three paradigms. Many of the existing researches on the usability of faceted search support this conclusion. Faceted search typically achieves higher precision and recall, and uses less time to retrieve the results [2, 6, 21, 32, 34, 41, 45-48], comparing to keyword search, especially in

the scenario where the users are unfamiliar with the topic being searched. Previous literature [21, 49-51] also pointed out that faceted search outperforms the monolithic taxonomy-based directory search in terms of classification effectiveness, and has better support for multi-criteria indexing and so on.

The major limitation of faceted search is that most faceted taxonomies in existing faceted search systems are still created manually by domain experts. It is time consuming and of high labor cost [52].

1.3 Faceted taxonomy of research work on faceted search

By analyzing the extensive faceted search research of the past decade, we developed a faceted taxonomy representation of these literatures, which is shown in Figure 2. The faceted taxonomy consists of four facets: facet models, key technologies, evaluation matrices and faceted search systems. When looking at "*facet models*" property, the top-level terms are "*theoretical basis*," "*model structure*", "*main terms*", "*interactivity*" and so on. The "*key technologies*" facet consists of five top-level terms: "*facet term extraction*", "*hierarchy construction*", "*compound term generation*", "*facet ranking*" and "*search results ranking*". They correspond to the five stages of faceted search. The "*evaluation matrices*" facet contains four terms: "*subjective metrics*", "*objective metrics*", "*relevance metrics*" and "*cost-based metrics*". The "*faceted search systems*" facet has top-level terms of "*data type*", "*facet type*," "*generation of faceted taxonomy*", "*integration with other search methods*" and so on, in accordance with the features and specifications associated with faceted search system.

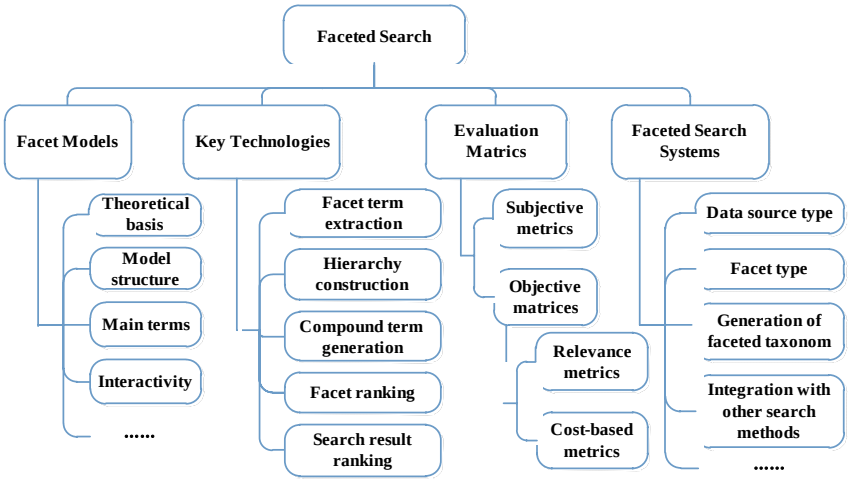

Figure 2. The faceted taxonomy of research work on faceted search

Under the guidance of the above faceted taxonomy, we surveyed the existing work on faceted search in detail. The rest of this paper is organized as follows. Section 2 reviews some well-known facet models. Section 3 proposes a general framework for faceted search system, and analyzes the key technologies of faceted search, including facet term extraction, hierarchy construction, compound term extraction and facet ranking. Section 4 mainly discusses two types of objective evaluation metrics for faceted search: relevant matrices and cost-based metrics. In Section 5, we compare the performance of several state-of-art faceted search systems. Finally, the conclusions and future work are presented in Section 6.

2 Facet Model

Facet model is the formal description of a faceted taxonomy and the facet-based navigation process. According to different modeling methodologies, facet models can largely be divided into three categories: set theory based, FCA (Formal Concept Analysis) [53] based, and lightweight ontology [54] based. In Table 2, we briefly compared the facet models mentioned above by theoretical basis, model structure and other key aspects.

Table 2 Comparison of existing facet models

Theoretical Basis	Model Name	Time	Model Structure	Main Concepts	Interactivity		Application
					Filtering	Ranking	
Null	HFC	2003	Flat /Hierarchical	Faceted metadata	Null	Null	Flamenco [21]
Set theory	Sacco's Model	2000	Tree	Terminology, Subsumption, Faceted taxonomy, Taxonomy-based source, etc.	Focus, Zoom points, Restriction	Null	dbWorld Xtended [58]
	FaSet	2009	Tree	Facet, Facet space, Multidimensional, Classification, etc.	Filtering computation	Focus similarity	Freeable [59]
	Li's Model	2010	DAG	Category hierarchy, Facet, Navigation path, etc.	Null	Facet ranking	Facetedpedia [9]
FCA	FKR	2000	Lattice	Facet, Interpretations, Units, Relations, etc.	Null	Null	FaIR[62]
Lightweight ontology	Faceted Lightweight Ontology	2009	Tree	Facet, lightweight classification ontology, etc.	Null	Null	Living Knowledge Project

Three typical set theory based models are discussed as follows:

The faceted taxonomy model proposed by Sacco et al. [55, 56] organizes the facet terms into a tree structure by means of hierarchical relation ("*is-a*" or "*part-of*"). This model provides the formal definitions of *terminology*, *subsumption*, *compound term*, *taxonomy*, *faceted taxonomy*, *interpretations* and *taxonomy-based source*. It also provides necessary simple and intuitive operations to support interactivity, such as *zoom-out* and *zoom-in* used for exploring the tree structure. In recent years, the development of Compound Term Composition Algebra (CTCA) [44] has significantly improved this model by providing a series of algebra operations. This model has been widely adopted by faceted search systems such as Mitos [57], dbWorld Xtended [58].

FaSet model proposed by Bonino et al. [59] focuses on using structured data in relational database. It provides the formal definitions of *facet*, *facet space*, *focus*, *multi-dimensional classification*, etc. In addition FaSet offers two search algorithms for faceted navigation and search results ranking respectively, which are implemented in SQL.

Li et al. [9] presented a facet model with a Directed Acyclic Graph (DAG) structure based on the set theory. It is used in Facetedpedia system. This model gives the formal definitions of *category hierarchy*, *facet*, *navigation path*, *faceted interface*, and provides the *facet ranking* operation based on the navigation cost and pairwise similarity among facets.

Faceted Knowledge Representation (FKR) model from Uta Priss [60] is a typical FCA-based facet model. FKR model gives the formal definitions of *unit*, *relation*, *facet*, *interpretation*, and organizes facet terms into a lattice structure. In contrast to the above set theory based models, FKR can only map data items to a single taxonomy and has no interactivity.

Faceted Lightweight Ontology proposed by Giunchiglia et al. [61] is a typical lightweight ontology. It has a rooted tree structure where each node is associated with a natural language label. The labels of nodes are organized according to certain predefined patterns that capture different aspects of items. This model gives the definitions of *category ontology*, *lightweight ontology*, and *faceted lightweight ontology*, but does not provide interactive operations.

In addition to the above models, there are a small number of facet models without explicit theoretical basis, such as Hierarchy Faceted Categories (HFC) model proposed by Yee et al [21].

It can be seen that the models based on set theory are capable of modeling facet ranking and faceted filtering, and are better in interactivity than other models.

3 Key Technologies

Through an empirical analysis of various faceted search systems such as Flamenco, Facetedpedia, mSpace [63], we propose a general faceted search framework, as illustrated in Figure 3. The framework consists of three modules: faceted taxonomy generation, query refining and result ranking. They cover four key technologies, facet term extraction, hierarchy construction, compound term generation and facet ranking.

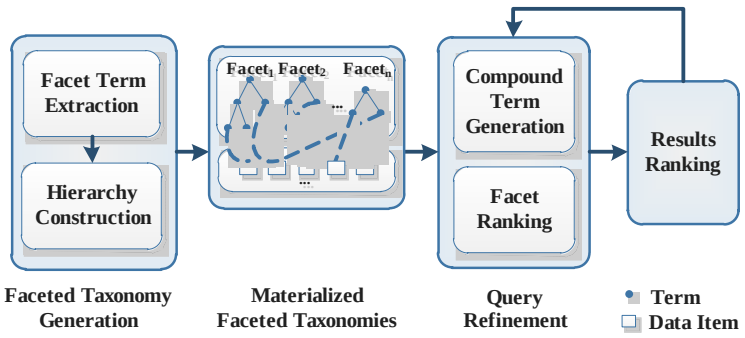

Figure 3. The general framework for faceted search system

The faceted taxonomy generation module constructs a faceted taxonomy from various text sources in an automatic or semi-automatic way. This module consists of two sub-modules: facet term extraction and hierarchy construction. The goal of the former is to automatically extract facet terms from structured or unstructured text data, such as XML and hypertext files; while the latter discovers the hierarchical relationships among facet terms. During the construction of faceted taxonomy, some algorithms can also establish the relations between the faceted taxonomy and the data items, and produce a materialized faceted taxonomy at the same time.

By means of a faceted taxonomy, the query refining module can facilitate users to select facets and refine search results in an iterative way. The core functionalities of this module include compound

term generation and facet ranking. The first function generates all the reasonable compound terms comprising one or more facet terms, in order to determine whether the compound term selected by a user is valid. The second function selects only a portion of all the facets to display in the user interface, when there are too many applicable facets and facet terms. The query refining process runs as follows. First, the facets in a faceted taxonomy are ranked, and some are selected to be showed in the user interface. The terms in the selected facet are employed by user to form compound terms. Secondly, the system validates the compound terms selected by the user, and shows only the data items associated with the valid compound terms. At the same time, the system updates the number of data items corresponding to facet terms in user interface and ranks the facet terms again for the next navigation activity. The iterations continue until expected results are found.

Search results ranking in the faceted search is similar to that in the traditional Information Retrieval domain. It has been extensively studied for years [64-66]. Thus we will skip it from the following sections.

3.1 Facet term extraction

Manual recognition of facet terms by domain experts is costly, inefficient and has poor scalability. A few researchers have conducted a preliminary study of automatic facet term extraction. Existing automatic extraction methods can be divided into three categories corresponding to the different data types: unstructured, semi-structured and structured.

Unstructured data refers to the information that does not adhere to a predefined data model. In facet term extraction, the most common form of unstructured data is the natural language text, which is always ambiguous and ill-formed. Since machine understanding of natural language text remains as an open topic, it is very hard to automatically extract facet terms by using machine learning techniques only. Current extraction methods mainly focus on making comprehensive use of the statistics of terms, linguistic features of the terms and external knowledge base. The typical methods are outlined as follows.

Stoica et al. [67] proposed Castanet algorithm to select facet terms based on term frequency distribution. The essence of this algorithm is selecting the terms having a frequency higher than a threshold as facet term candidates for subsequent processing. This algorithm can be easily implemented and extended to different domains since only term frequency is employed.

Anick and Tipirneni [68] proposed a facet term extraction algorithm based on the lexical dispersion of words in text. Lexical dispersion of a word is the number of different compounds that contain this word within a document collection. The algorithm consists of two stages. In the indexing stage documents are parsed so the lexical compounds can be extracted. In the querying stage the compounds appearing in the top n documents of a ranked result list are used to compute the lexical dispersion of each term occurring within these compounds. These terms are then sorted by their dispersions and the top m terms are selected as candidate terms for subsequent hierarchy construction. The disadvantage of this algorithm is that the extraction of facet terms depends on the specific lexical structure and therefore can be hardly extended to new domains.

Ling et al. [69] proposed a two-stage probabilistic method to extract facet terms based on topic model. Given the original keywords from a user, this method first applies a bootstrapping algorithm to the document collection to get more correlated terms. Probabilistic mixture models are applied to these

expanded terms to estimate the term distribution of every facet. This is done by simultaneously fitting the topic model to the data set and restraining the model so that it is close to the specified definition from the user. The basic idea behind the processes is to guide the topic model with user-defined keywords.

Dakka and Ipeirotis proposed an unsupervised automatic facet extraction algorithm using external resources [51]. This algorithm first identifies the facet term candidates in each document by using third-party term extraction services or algorithms, such as LingPipe [70], Yahoo Term Extraction [71], Wikipedia, or Taxonomy Warehouse [72]. Then, each candidate is expanded with "context" phrases appearing in external resources by querying WordNet, Wikipedia, and other online dictionaries. This step produces the latent facet terms in the expanded term set, which do not explicitly appear in the documents. Finally, the term distributions in the original term set and the expanded term set were compared to identify the terms that can be used to construct browsing facets. This algorithm has good flexibility and extensibility. However the quality of the extracted facets heavily depends on the quality of the external resources and term extractor.

The semi-structured data does not conform to an explicit data schema; however, it generally contains tags or other markers to separate semantically related elements. Examples of the semi-structured data include HTML pages and the pages annotated by Resource Description Framework (RDF). Semi-structured data has an implicit formal structure, which can be exploited to improve the quality of facet term extraction. For example, the hyperlinks of web pages can be used to evaluate the importance of facet terms. The typical extraction methods in this subfield are described briefly as follows.

Li et al. [9] developed a system named Facetedpedia that exploits internal hyperlinks of Wikipedia and folksonomy for automatic extraction of facet terms. Facetedpedia regards titles of articles as facet terms, and constructs the taxonomy of articles that is hyperlinked from user keywords query results, based on the Wikipedia category system. Oren et al. [41] proposed a facet term recognition method dedicated to the semi-structured data in semantic web. This method was implemented by dynamically constructing a faceted navigation tree based on RDF graph.

Structured data has an explicit data model or schema, such as data stored in a relational database. For structured data, the core task of facet term extraction is to select facet terms from attributes of database. Roy et al. [73] presented such a method. At every step, a user is asked one or more questions about the different facet terms, and the most promising set of facet terms is identified based on the user's response. Zhao et al. [74] implemented TEXplorer which selects facet terms from the attributes by measuring the relevance between keywords and documents.

3.2 Hierarchy construction

Hierarchy construction aims at discovering the "*is-a*" and "*part-of*" relationships among the facet terms extracted from text corpus. Currently, clustering-based and pattern-based methods are quite popular.

The clustering-based methods exploit the semantic similarity or semantic distance between concepts to induce hierarchical relations among them. In these methods, a cluster can be considered as a facet term. The hierarchical structure of the clusters corresponds to the relations among the facet terms.

The hierarchical relation extraction algorithm proposed by Zeng et al. [75] takes the search keywords as the root node. The algorithm organizes the search results into the branches of the root node and assigns each branch a facet term. An Support Vector Machine regression model is learned from human labeled training data to search facet terms in the search results. Each search result is given relevant facet terms to form multiple candidate groups. The final groups are generated by merging the initial candidate groups. This method cannot generate multidimensional taxonomies. Also the extracted relations may be neither "*is-a*" nor "*part-of*".

Dou et al. [76] developed QDMiner system to automatically extract facet hierarchies for keywords by aggregating frequent lists from free text, HTML tags, and repeated regions within top search results. The process is described as: 1) When a user issues a query, QDMiner retrieves the top-k results from a search engine to form a set *R*. 2) QDMiner extracts and weights several types of lists from each document of *R*. 3) QDMiner groups similar lists together to form a facet by a modified quality threshold clustering algorithm. 4) QDMiner evaluates and ranks facets and facet items based on their importance.

Chen and Li [77] presented a method of automatic term hierarchies acquisition based on subsumption estimation and spectral clustering. First, each term is considered as a vertex in an undirected weighted graph. The problem of hierarchical relation construction is then modeled as a modified graph-partitioning problem and is solved by spectral clustering methods. Subsumption estimation is introduced to guide the spectral clustering process. As a result, a modified spectral graph partitioning algorithm was developed to accurately depict the hyponym information about facet terms. This method can extract facet terms based on compound words such as "probabilistic clustering," but the hierarchical taxonomies may be significantly different from the manually obtained taxonomies.

The methods based on hierarchical concept clustering do not need training data and hence can easily be applied to a wide range of domains. The disadvantages of these methods are, firstly, it may be hard to give the produced cluster a category name. Secondly, the hierarchical structure is not in accord with the cognitive structure that most people know.

Pattern-based methods primarily exploit specific patterns to construct the hierarchical structure. The applicable patterns include the co-occurrence of facet terms, the existing hierarchical relations in a semantic database such as WordNet, FreeBase and FrameNet [78].

The subsumption algorithm proposed by Sanderson and Croft [79] can automatically extract hierarchical relation from a set of documents using term co-occurrence. This algorithm assumes that a term pair (*x, y*) meeting the restrictions of $P(x|y) > 0.8$ and $P(y|x) < 1$, has hierarchical relation. Then Dakka [80] optimized the algorithm and reduced the time complexity from $O(n^2)$ to $O(n^2/m)$, where *n* is the number of terms in text and *m* is the number of facet terms. This algorithm only depends on statistical features of terms to determine the hierarchical relation. This may result in low precision.

Castanet algorithm presented by Stoica and Hearst [67] can automatically generate Hierarchical Structure of faceted metadata from textual descriptions of data items by exploiting the "*is-a*" relationship in WordNet. The algorithm selects candidate terms from textual descriptions of data items. Then the terms having only one sense are used to build the core tree based on the "*is-a*" relation defined by WordNet. Finally the ambiguous terms and low-frequency terms are added to extend the core tree. Through the above steps, a faceted taxonomy is constructed automatically. This algorithm solely depends on WordNet to obtain the "*is-a*" relationship and therefore has limited scalability.

Deep Classifier proposed by Xing et al. [81] is a novel algorithm that groups search results into detailed hierarchical categories by pruning online web directories. The web directories, such as Yahoo!Directory and the Open Directory Project, are always too complex to construct an appropriate taxonomy directly. Therefore, this algorithm first searches for the web directory by using keywords from a user query. Then a hierarchical structure related to user keywords is generated from the search results. Compared to the online web directories, the taxonomic structure contains fewer nodes and is closely related to the user keywords.

Hierarchical structure constructed by the pattern-based methods is more conforming to human cognitive structure, compared to those constructed by the clustering-based methods. However, pattern-based methods rely too much on the external semantic resources. If some facet terms do not appear in these semantic databases or taxonomy categories, it is hard to discover the hierarchical relations of these terms.

The relationships constructed by above methods are mostly of the binary type. Two terms either have a hierarchical relationship or they do not. In real world, however, the relationship between two terms is generally obscure and can be measured as a value in the range of [0, 1]. A preliminary study of fuzzy relation construction has been conducted, such as the literature [82, 83]; and more investigations should be launched by more researchers and research groups.

3.3 Compound term generation

A compound term is valid if it can index one or more data items. In general, most valid compound terms are manually built by domain experts. When there are a great number of facets or facet terms, it is necessary to automatically generate valid compound terms.

The key finding in compound term generation research is the Compound Term Composition Algebra (CTCA) [25, 44] proposed by Tzitzikas et al. It can generate all valid compound terms of a faceted taxonomy effectively. The expression of CTCA consists of a series of pre-defined valid compound terms and invalid compound terms. All the terms are linked together by four basic operations: plus-product, minus-product, plus-self-product and minus-self-product. The algebra validates a compound term by recursively analyzing the syntax tree of the expressions affiliated to the faceted taxonomy. The operations of plus-product and plus-self-product are used to judge whether or not a compound term is the parent node of some valid predefined compound terms. The minus-product and minus-self-product operations are used to judge whether a compound term is the child node of an invalid predefined compound term.

Since CTCA only needs to store the two kinds of expressions which are formed respectively by valid or invalid compound terms; it can validate a compound term in shorter running time with less memory consumption.

3.4 Facet ranking

When there are too many facets or facet terms, or the user interface is too small to display them all, only some of these facets or terms should be displayed. This requires the facets and terms to be ranked so that the most important ones are picked. Our survey identified two major types of facet ranking methods: the independent facets (attributes) based and the correlated facets based.

The independent facet based ranking methods mainly depend on classification capacity of facet. The representative researches are briefly described as follows:

Oren et al. [41] exploited the predicate balance, object cardinality and predicate frequency to rank facets. The predicate balance is referred to as the balance of the faceted navigation tree composed of faceted attributes (terms). The more balanced this navigation tree is, the higher the navigation efficiency is. The object cardinality is the number of values that can be assigned to the faceted attribute. The smaller this cardinality is, the easier it is for user to select a suitable value. The predicate frequency indicates the classification capacity of a facet. If the predicate frequency of a faceted attribute is low, when a user selects a value of this attribute, only a small number of data items are affected.

Roy et al. [84] considered the navigation process in a structured database as the decision process for constructing minimum cost tree (the average path length between all leaf and the root of the tree is minimized). The attribute represented by the root node of the decision tree is the facet with the best classification capacity. The classification capacity of a facet is defined by the formula $Indg(A_l, D) = \sum_{1 \leqslant q \leqslant |Dom_l|} |D_{x_q}|(|D_{x_q}| - 1)/2$, where A_l is the attributes (terms) of one facet, *D* is the data set, Dom_l is the domain of attribute A_l, x_q is the *q*-th value assigned to A_l, and D_{x_q} is the subset of *D*, in which value of A_l is x_q. The facet with the minimum value of *Indg()* is selected as the optimal facet. By continuing calling *indg()*, the desired *k* facets will be obtained.

Ranking methods based on the correlated facets employ the overlap, mutual information and conditional probability of facets or facet terms to select or rank facets. Although more reasonable, these methods tend to increase the complexity of facet selection and ranking problem. Noticable works are briefly described as follows:

Li et al. [9] ranked individual facet hierarchies by measuring user's navigational cost. They also ranked multiple facets by using both their average navigational costs and average pairwise similarities. The navigational cost is defined as the logarithm of the length of the navigation path. The average pairwise similarity of *k* overlapped facets is formally defined as $sim(\{F_1, ..., F_k\}) = \sum_{1 \leq i \leq j \leq k} 2 \times sim(F_i, F_j)/k(k-1)$, where $sim(F_i, F_j)$ between F_i and F_j is defined as the overlapped items between the facet F_i and F_j.

The method proposed by Zwol et al. [3] ranked the candidate facets based on the statistical analysis of query terms and query sessions derived from the image search logs. The first step is to calculate all possible co-occurring objects for a query event. Then a series of metrics are computed, including atomic metrics (probability, entropy), symmetric metrics (cosine similarity, joint probability) and asymmetric metrics (conditional probability, K-L divergence).

Yamamoto et al. [85] applied Co-HITS framework to rank the facets extracted from community QA corpus. Co-HITS is a generalized form of the Hyperlink-Induced Topic Search (HITS) algorithm [86], which stochastically calculates the importance of the nodes in a bipartite graph $G = (E \cup F, \mathcal{E})$, where the vertex set $E = \{e_i\}_n$ and $F = \{f_j\}_m$ correspond to the entities and the facets respectively. The edge set $\mathcal{E}$ contains edges between vertices in E and F.

The above literatures rank facets or facet terms. However ranking the faceted values of the individual facet terms is also necessary in some cases. At each navigation step, when too many results

queried by compound term lead to too many faceted values, some rules or criteria should be applied to give higher priority to only some of the results. The representative works are listed as follows.

Kashyap [6] proposed a faceted navigation model that considers a candidate facet having minimum navigation cost as the desired facet. This model defines three types of operations during navigation: SHOWRESULT, REFINE, and EXPAND. Kashyap proved that the calculation of the minimum navigation cost is an NP-hard problem and therefore provided two approximation algorithms.

Dash et al. [42] exploited random sampling to measure user interests in faceted values, based on the frequency of facets in text. Given a faceted value *f*, *p*-value is used to estimate the user interests. The *p*-value is defined as $1-\sum_{k=0}^{q-1}\binom{r}{k}\binom{R-r}{Q-k}/\binom{R}{Q}$, where *R* is the number of documents in the data set, *r* is the number of documents related to *f*, *Q* is the number of query results, *q* is the number of documents related to *f*. The lower the *p*-value is, the higher the interest is. For multiple faceted values, this method considers the *p*-values of the *k* most interesting values in every facet, the accumulated value $\sum_{i=1}^{k}\omega_i s_i$ is used to estimate the user's interests, where ω_i is the weight of the attribute value and $s_i=-\log p_i$. The larger the accumulated value is, the more the total interests are.

Koren et al. [31] developed a facets-based general probabilistic framework to build a document model. They also proposed a uniform method to solve the facet ranking problem. The ranking criteria that this method uses are the basic relevance between facets and individual users, and the collaborative relevance between facets and multiple users. Suppose there exists a set of training documents χ_u, the number of training documents related to a user *u* is $\chi_{u,rel}$. The probability of the relevance between the facets and the user is defined as $P(x_k|rel,u)=1/|\chi_u|\times\sum_{x\in\chi_{u,rel}}x_k$, where x_k represents the document *x* containing the *k*-th pair of facet term and value. The bigger $P(x_k|rel,u)$ is, the higher the relevance is.

4 Evaluation Metrics

Metrics used to evaluate faceted search can be divided into subjective metrics and objective metrics. The former can be used to assess users' subjective judgment of search quality. For example, the subjective metrics proposed by Yee [21] and Uddin [87] include "*easy to use*," "*easy to browse*," "*simplicity*" and "*flexibility*." Bartolini [49] and Smith [88] adopted the average satisfaction rate as the overall search quality metric. The disadvantage of subjective metrics is that the evaluation results are likely to be affected by users' mental state, cognitive level and so on. Objective metrics evaluate search results and search process by adopting standard benchmark. Here, we mainly focus on two types of objective metrics: relevance metrics and cost-based metrics.

4.1. Relevance metrics

Relevance metrics are used to evaluate how relevant a search result is regarding a given query. In faceted search, the matching between data items and facet terms in many cases are predetermined. Only a small number of faceted search systems support automatic classification of search results based on facet terms [52]. Therefore, the relevance metrics of faceted search results always are high.

At present, a series of metrics have been proposed by information retrieval community to describe the binary relevance and graded relevance. These metrics have been applied to faceted search without modification [89].

Commonly used binary relevance metrics include precision, recall, F-measure, E-measure and their macro and micro forms. For example, Gomadam [90] adopted precision and recall to measure search quality. Xing et al. [81] used micro-F1, macro-precision, macro-recall, macro-F1 to evaluate the results of their Deep Classifier faceted search system.

Graded relevance metrics mainly include r-precision (rPrec), normalized Discounted Cumulative Gain (nDCG) [91], Rank biased Precision (RBP) [92], Mean Average Precision (MAP) [93], Mean Reciprocal Rank (MRR) [94] and Binary Preference (BPref) [95]. Macdonald et al. [96] employed MAP, rPrec and bPref to evaluate faceted search in blogs. Zhang et al. [8] adopted MAP, R@1000 and other metrics to evaluate the faceted search on OHSUMED and RCV1 datasets. Pound et al. [97] exploited nDCG to rank the output of their facet discovery algorithm.

4.2. Cost-based metrics

The cost-based metrics mainly look at run-time overhead and memory usage. Karlson [30] employed the completion time of retrieval tasks as the evaluation metrics. He compared the efficiency of faceted search and keyword search on mobile devices. Uddin et al. [87] used mean standard variance and mean difference of completion time as evaluation metrics to measure the performance of faceted search and keyword search. The completion time of retrieval tasks usually follows a positively-skewed distribution. In order to make the value of the metric follow a normal distribution, Xu and Dumais et al. [2, 98] adopted the logarithm of the average completion time as the evaluation metrics.

Dash et al. [42] employed two cost-based metrics: the time spent on calculating the number of attribute-value pairs of facet terms, and the memory usage in index storing process.

5 Faceted Search Systems

After the first well-known faceted search system Flamenco, a series of faceted search systems have been developed and applied to a variety of domains. In an effort to provide an overall picture of the faceted search systems, we compared twenty existing systems in terms of data source, facet term extraction, hierarchy construction, facet ranking and so on in Table 3.

Table 3 Comparison of the existing faceted search systems

Name	Time	Data source	Facet term extraction	Hierarchy construction	Facet ranking	Other search paradigms integrated
Flamenco	2003	Relational database	Manually selecting attributes from database	Generated manually	None	Keyword search
DynaCet	2009	Relational database	Automatically selecting attributes from database	Based on users' minimum cost	Navigation cost based	Form-based search
TEXplorer	2011	Text database	Automatically selecting attributes from database	Based on users' selection	Significance measure	Form-based search

FACeTOR	2010	Relational database	Automatically selecting attributes from database	Not mentioned	Navigation cost based	Keyword search
IBM PES	2011	Text database	Automatically selecting attributes from database	Not mentioned	None	Keyword search
Facetedpedia	2010	Wikipedia pages	Extracting the titles of Wikipedia pages	Based on Wikipedia category system	Average similarity of *k*-Facet	Keyword search
Mitos	2008	Web pages	Clustering web pages	Organized by users	None	Form-based search
Relational Browser++	2005	Web pages	Clustering web pages	Generated manually	None	Form-based search
Blognoon	2011	Web pages	Automatically extracting from blog posts	Clustering blog posts	Relevance to a search query	Keyword search
MediaFaces	2010	Semi-structured data	Automatically selecting from annotation information	Facet extraction	Using the latest query logs	Keyword search
mSpace	2005	RDF data	Automatically selecting attributes of RDF data	Organized by users	None	Keyword search
/facet	2006	RDF data	Automatically selecting attributes of RDF data	Organized by users	None	Keyword search
Longwell	2005	RDF data	Automatically selecting attributes of RDF data	Organized by users	None	Keyword search
Tabulator	2006	RDF data	Attributes of RDF data	Not mentioned	None	None
Parallax and Companion	2009	RDF data	Attributes of RDF data	Not mentioned	None	Keyword search
Stuff I've seen	2003	Unstructured local documents	Not mentioned	Organized by users	None	Keyword search
DocuBrowse	2010	Unstructured enterprise documents	Manual	Organized by users	None	Form-based search
FacetLens	2009	Unstructured academic articles	Not mentioned	Not mentioned	None	Keyword search
CiteSeerX	2008	Unstructured academic papers	Using metadata of documents	Not mentioned	The impact of citations	Keyword search Form-based search
Apache Solr	2004	Unstructured documents	Using metadata of documents	Not mentioned	None	Keyword search

Flamenco [21] of UC Berkeley, DynaCet [73] of University of Texas at Arlington (UTA), TEXplorer [74], FACeTOR [6] of University at Buffalo Buffalo and IBM PES (Patient Empowerment System) [99] can build a faceted taxonomy, either manually or automatically, from structured data such as rational databases, text databases, which have structured attributes available.

Facetedpedia [9] of of UTA extracts facet terms from the titles of Wikipedia pages, and constructs a faceted taxonomy based on Wikipedia category system. Mitos [57] of Forth and Relational Browser++ [100] of University of North Carolina generates main facets by clustering the crawled pages from websites, but the faceted taxonomy was built manually. Blognoon [101] extracts facet terms by using a key term extraction algorithm, and constructs the facets hierarchy by building blog posts clusters.

MediaFaces [3] of Yahoo! can extract a faceted taxonomy from the semi-structured annotation data and rank facets based on user query logs. mSpace [63] of Southampton, /facet [102] of CWI institute, Longwell [103] of MIT Tabulator [104], and *Parallax and Companion* [105] implement faceted search and navigation for semi-structured Semantic Web data (RDF data). mSpace, /facet and Longwell support automated faceted taxonomy construction, and allow users to modify the faceted taxonomy if necessary. *Parallax and Companion* supports set-based browsing. It allows users to efficiently browse through graph-based data by moving from a dataset to another related dataset.

Stuff I've Seen [106] of Microsoft, DocuBrowse [23] of TAMU and other systems implemented various document management systems to support the faceted navigation. Stuff I've Seen implemented uniform file management system, with which users can organize all types of local resources based on a faceted taxonomy. DocuBrowse can recognize certain types of documents, and can automatically recommend some documents to users based on document similarity and search history.

Relational Browser++ [100] of UNC and FacetLens [107] not only support faceted navigation, but also have certain data analyzing functionalities. They can represent the distribution of different facet values by a histogram.

In addition, there are some widely used open-source faceted search systems. CiteSeerX [108] is built with a new open source infrastructure, SeerSuite, and indexes more than two million documents. Apache Solr [109, 110] is an open source search platform supporting faceted search and full-text search. It powers the search functionality of public sites such as CNET, Zappos, and Netflix, as well as countless other government and corporate intranet sites. The source code and demos of another open source system, Flamenco, can be accessed freely at its project website [111].

6 Future Research Directions

Although faceted search has been under intensive study, there are still a number of problems to be solved. Based on our survey, we outline some possible future research directions of faceted search research.

a) Automatic faceted taxonomy construction

The faceted taxonomies of most existing faceted search systems are manually created and maintained, which is time consuming and labor intensive. Currently, although preliminary study of the automatic faceted taxonomy construction has been carried out [51, 67], the existing methods are still far from practical.

The future research on this subject should focus on the automatic extraction of the facet terms and their hierarchical relations. In the ontology learning research community, extensive research on extracting domain terms and their relations has been carried out. Nevertheless, these methods cannot be applied to the extraction of facet terms and their relations, because (1) Facet terms and domain terms are substantially different from each other, in that the domain terms are always domain-specific,

while a portion of the facet terms can be domain-independent. (2) Domain terms are usually distributed in data objects (for example, domain text). However, the top-level facet terms rarely appear in data objects [51]. For example in Figure 1, the content of literature generally does not include the top level facet terms such as "*Subject Areas*," "*Computer Science*". (3) Constrained by the orthogonality of facet taxonomy, an individual facet term only belongs to one particular facet. The facet terms of different facets have no hierarchical relations.

b) The faceted classification and faceted navigation for complex, open data

Up till now, faceted search mainly focuses on closed datasets. The data types of these data sets are mostly structured or semi-structured. With the rapid growth of the web applications, more and more unstructured data or complex structured data are produced. Therefore, novel faceted classification and faceted navigation methods for these data should be extensively studied.

c) Faceted interface and visualization

The visual clutter and change blindness are two unsolved problems of visualization. The causes of the visual clutter include excessive number of facet terms and the limited size of the user interface. Therefore, adaptively displaying, hiding, expanding and folding of facet and facet terms need to be investigated in the near future. Change blindness is a perceptual phenomenon where an observer fails to notice visual changes [112]. In faceted search, the change blindness refers to the scenario that a sudden disappearing of data objects during a navigation process. This often the users [113]. To solve this problem, the animated transitions of faceted search should be able to help users perceive changes in the interface.

In addition, representing the search results in a form of a multi-dimensional cube is also a future research direction. Compared to the data presentation in a one-dimensional cube, which is widely adopted by most existing faceted search systems, a multi-dimensional cube not only improves the search efficiency, but can also easily support the integration of data mining, OLAP Cube and so on [22, 114].

d) Relevance evaluation of faceted search results

Currently, the metrics used to evaluate information retrieval systems are adopted by faceted search community without any modification. They are not very applicable in some real-world situations. The reasons are: (1) These metrics are mainly designed to evaluate the search results organized in a "one-result-list-only" way. They may not be suitable for measuring the results of faceted search [115]. (2) Most evaluation metrics for information retrieval systems follow the assumption of independent relevance [116]. While in faceted search, new search results are always the subset of the previous one, which violates the assumption of independent relevance. Therefore, new evaluation metrics considering category-based search results representation and dependent relevance should be studied.

7 Conclusions

Faceted search is a technique of accessing a large collection of information that is represented by a faceted taxonomy. It enables users to select facets and facet terms to refine the search results in an iterative way. Extensive research has been done in this domain during the last decade. This paper summarized the published literatures, and proposed a faceted taxonomy of research work on faceted search. On the foundation of the taxonomy, three types of facet models, which are based on the set theory, FCA and ontology respectively, are compared in various aspects. We then propose a general faceted search framework; four key technologies in the framework, namely facet terms extraction,

hierarchy construction, compound term extraction and facet ranking, are reviewed in detail. After that, the relevance metrics and cost-based metrics for faceted search evaluation are explained. Finally, a comparative study on the existing faceted search systems is conducted.

Although a variety of faceted search systems have been applied to many domains, there are still several open questions to be solved. This paper points out four future research directions, including automatic faceted taxonomy construction, faceted interface and visualization, relevance evaluation of faceted search results, and the faceted classification and navigation of complex data and open data.

Acknowledgements

The research was supported in part by the National High-Tech R&D Program of China under Grant No.2012AA011003, the National Science Foundation of China under Grant Nos. 60825202, 61173112, 60921003,61202184, Cheung Kong Scholars Program, Key Projects in the National Science & Technology Pillar Program under Grant No. 2011BAK08B02, 2011BAK08B05. The authors are grateful to the anonymous reviewers for their comments, which greatly improved the quality of the paper.

References

1. Furnas, G.W., et al., The Vocabulary Problem in Human System Communication. Communications of the ACM, 1987. 30(11): p. 964-971.
2. Dumais, S., Cutrell, E., and Chen, H., Optimizing search by showing results in context, in Proceedings of the SIGCHI conference on Human factors in computing systems. 2001, ACM: Seattle, Washington, United States. p. 277-284.
3. Zwol, R.v., et al., Faceted exploration of image search results, in Proceedings of the 19th international conference on World Wide Web. 2010, ACM: Raleigh, North Carolina, USA. p. 961-970.
4. Tunkelang, D., Dynamic Category Sets: An Approach for Faceted Search, in Proceedings of the ACM SIGIR'06 Workshop on Faceted Search. 2006: Seattle, WA, USA.
5. Bergamaschi, S., Guerra, F., and Leiba, B., Information Overload Internet Computing, 2010. 14(6): p. 10-13.
6. Kashyap, A., Hristidis, V., and Petropoulos, M., FACeTOR: Cost-Driven Exploration of Faceted Query Results, in ACM Conference on Information and Knowledge Management (CIKM). 2010: Toronto, Ontario, Canada.
7. Marshall, P., Herman, S., and Rajan, S., In search of more meaningful search. Serials Review, 2006. 32(3): p. 172-180.
8. Zhang, L. and Zhang, Y., Interactive retrieval based on faceted feedback, in Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010, ACM: Geneva, Switzerland. p. 363-370.
9. Li, C., et al., Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia, in Proceedings of the 19th international conference on World Wide Web. 2010, ACM: Raleigh, North Carolina, USA. p. 651-660.
10. Hearst, M., UIs for Faceted Navigation: Recent Advances and Remaining Open Problems, in HCIR08 Second Workshop on Human-Computer Interaction and Information Retrieval. 2008: Redmond, WA.
11. eBay.Available from: http://www.ebay.com/.
12. Amazon.com. Available from: http://www.amazon.com.

13. Diederich, J. and Balke, W., FecetedDBLP - Navigational Access for Digital Libraries. Bulletin of the IEEE Technical Committee on Digital Libraries (TCDL), 2008. 4(1).
14. IEEE Xplore - Home. Available from: http://ieeexplore.ieee.org/Xplore/dynhome.jsp?tag=1.
15. ISI Web of Knowledge. Available from: http://isiknowledge.com.
16. The Open Video Project. Available from: http://www.open-video.org/.
17. Google Images. Available from: http://images.google.com/.
18. Ranganathan, S.R., Elements of library classification (1st ed). 1991, Bombay, New York: South Asia Books. 168 p.
19. Prietodiaz, R., Implementing Faceted Classification for Software Reuse. Communications of the ACM, 1991. 34(5): p. 88-97.
20. Spiteri, L., A Simplified Model for Facet Analysis. Canadian Journal of Information and Library Science, 1998. 23(1-2): p. 1-30.
21. Yee, K.P., et al., Faceted metadata for image search and browsing, in Proceedings of the SIGCHI conference on Human factors in computing systems. 2003, ACM: Ft. Lauderdale, Florida, USA. p. 401-408.
22. Ben-Yitzhak, O., et al., Beyond basic faceted search, in Proceedings of the international conference on Web search and web data mining. 2008, ACM: Palo Alto, California, USA. p. 33-44.
23. Girgensohn, A., et al., DocuBrowse: faceted searching, browsing, and recommendations in an enterprise context, in Proceeding of the 14th international conference on Intelligent user interfaces. 2010, ACM: Hong Kong, China. p. 189-198.
24. Uddin, M.N. and Janecek, P., The implementation of faceted classification in web site searching and browsing. Online Information Review, 2007. 31(2): p. 218-233.
25. Tzitzikas, Y., Evolution of faceted taxonomies and CTCA expressions. Knowledge and Information Systems, 2007. 13(3): p. 337-365.
26. Jethava, V., et al., Scalable multi-dimensional user intent identification using tree structured distributions, in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011, ACM: Beijing, China. p. 395-404.
27. Fafalios, P., Kitsos, I., and Tzitzikas, Y., Scalable, flexible and generic instant overview search, in Proceedings of the 21st international conference companion on World Wide Web. 2012, ACM: Lyon, France. p. 333-336.
28. Taylor, A.G., Introduction to Cataloging and Classification (8th ed). 1992, Englewood, Colorado: Libraries Unlimited.
29. Dachselt, R., Frisch, M., and Weiland, M., FacetZoom: a continuous multi-scale widget for navigating hierarchical metadata, in Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. 2008, ACM: Florence, Italy. p. 1353-1356.
30. Karlson, A.K., et al., FaThumb: a Facet-based Interface for Mobile Search, in Proceedings of the SIGCHI conference on Human Factors in computing systems. 2006, ACM: Montréal, Québec, Canada. p. 711-720.
31. Koren, J., Zhang, Y., and Liu, X., Personalized interactive faceted search, in Proceeding of the 17th international conference on World Wide Web. 2008, ACM: Beijing, China. p. 477-486.
32. Hearst, M., et al., Finding the flow in web site search. Communications of the ACM, 2002. 45(9): p. 42-49.
33. Tunkelang, D., faceted search, G. Marchionini, Editor. 2009, Morgan & Claypool Publishers.
34. Sacco, G.M., Research results in dynamic taxonomy and faceted search systems, in the 18th International Conference on Database and Expert Systems Applications(DEXA). 2007: Torino, Italy p. 201-206, 862.

35. Allard, P. and Ferre, S., Dynamic Taxonomies for the Semantic Web, in Proceedings of the 19th International Conference on Database and Expert Systems Application. 2008, IEEE Computer Society: Turin, Italy. p. 382-386.
36. nsf.gov - Advanced Funding Search - US National Science Foundation (NSF). Available from: http://www.nsf.gov/funding/advanced_funding_search.jsp.
37. Hotelbook.com | Hotel Reservations | Find and Book Hotels with hotelbook.com. Available from: http://www.hotelbook.com/en/.
38. Yahoo! Directory. Available from: http://dir.yahoo.com/.
39. Open Directory Project. Available from: http://www.dmoz.org/.
40. Quintarelli, E., Resmini, A., and Rosati, L., FaceTag: integrating bottom-up and top-down classification in a social tagging system, in Proceedings of the 8th Information Architecture Summit. 2007: Las Vegas, Nevada, United States.
41. Oren, E., Delbru, R., and Decker, S., Extending faceted navigation for RDF data, in Proceedings of the 5th International Semantic Web Conference (ISWC). 2006. p. 559-572, 1001.
42. Dash, D., et al., Dynamic faceted search for discovery-driven analysis, in Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM). 2008, ACM: Napa Valley, California, USA. p. 3-12.
43. Clarkson, E.C., Navathe, S.B., and Foley, J.D., Generalized formal models for faceted user interfaces, in Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. 2009, ACM: Austin, TX, USA. p. 125-134.
44. Tzitzikas, Y., Analyti, A., and Spyratos, N., Compound Term Composition Algebra: The semantics. LNCS Journal on Data Semantics 2005. 2: p. 58-84.
45. English, J., et al., Hierarchical faceted metadata in site search interfaces, in CHI '02 extended abstracts on Human factors in computing systems. 2002, ACM: Minneapolis, Minnesota, USA. p. 628-639.
46. Kules, B., et al., What do exploratory searchers look at in a faceted search interface?, in Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. 2009, ACM: Austin, TX, USA. p. 313-322.
47. Kaki, M., Findex: search result categories help users when document ranking fails, in Proceedings of the SIGCHI conference on Human factors in computing systems. 2005, ACM: Portland, Oregon, USA. p. 131-140.
48. Yogev, S., et al., Towards expressive exploratory search over entity-relationship data, in Proceedings of the 21st international conference companion on World Wide Web. 2012, ACM: Lyon, France. p. 83-92.
49. Bartolini, I., A Multi-faceted Browsing Interface for Digital Photo Collections, in Proceedings of the 2009 Seventh International Workshop on Content-Based Multimedia Indexing. 2009, IEEE Computer Society. p. 237-242.
50. Fujisawa, S. and Andres, F., Multi-facet Category for Cultural Digital Resources, in Proceedings of the 21st International Conference on Data Engineering Workshops. 2005, IEEE Computer Society. p. 1227.
51. Dakka, W. and Ipeirotis, P.G., Automatic extraction of useful facet hierarchies from text databases, in 2008 IEEE 24th International Conference on Data Engineering. 2008. p. 466-475, 1631.
52. Hearst, M.A., Clustering versus faceted categories for information exploration. Communications of the ACM, 2006. 49(4): p. 59-61.
53. Wille, R., Restructuring lattice theory: an approach based on hierarchies of concepts, in Rival, I. (ed.): Ordered Sets. 1982, Boston. p. 445-470.

54. Giunchiglia, F., Marchese, M., and Zaihrayeu, I., Encoding Classifications into Lightweight Ontologies. Journal on Data Semantics, 2007. 8: p. 57-81.
55. Sacco, G.M., Dynamic taxonomies: a model for large information bases. IEEE Transactions on Knowledge and Data Engineering, 2000. 12(3): p. 468-479.
56. Sacco, G.M. and Tzitzikas, Y., Dynamic taxonomies and faceted search: theory, practice, and experience. The information retrieval series. 2009, Dordrecht, Netherlands; New York: Springer. 340.
57. Tzitzikas, Y., Armenatzoglou, N., and Papadakos, P., FleXplorer: A Framework for Providing Faceted and Dynamic Taxonomy-Based Information Exploration, in Proceedings of the 2008 19th International Conference on Database and Expert Systems Application. 2008, IEEE Computer Society. p. 392-396.
58. Sacco, G.M., DBWorld Xtended: Semantic Dissemination of Information through Dynamic Taxonomies Proceedings of I-KNOW 2005.
59. Bonino, D., Corno, F., and Farinetti, L., FaSet: A Set Theory Model for Faceted Search, in Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01. 2009, IEEE Computer Society. p. 474-481.
60. Priss, U., Description Logic and Faceted Knowledge Representation, in Proceedings of the 1999 International Workshop on Description Logics (DL'99). 1999: Linköping, Sweden.
61. Giunchiglia, F., Dutta, B., and Maltese, V., Faceted Lightweight Ontologies, in Conceptual Modeling: Foundations and Applications, A. Borgida, et al., Editors. 2009, Springer Berlin / Heidelberg. p. 36-51.
62. Priss, U., Faceted Information Representation. in Proceedings of the 8th International Conference on Conceptual Structures, 2000: p. 84-94.
63. Schraefel, M.C., et al., MSPACE: Improving information access to multimedia domains with MultiModal Exploratory Search. Communications of the ACM, 2006. 49(4): p. 47-49.
64. Haveliwala, T.H., Topic-sensitive PageRank: a Context-sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, 2003. 15(4): p. 784-796.
65. Liu, T.-Y., Learning to Rank for Information Retrieval. Foundations and Trends in Information Retrieval, 2009. 3(3): p. 225-331.
66. Ruthven, I. and Lalmas, M., A survey on the use of relevance feedback for information access systems. Knowl. Eng. Rev., 2003. 18(2): p. 95-145.
67. Stoica, E., Hearst, M.A., and Richardson, M., Automating Creation of Hierarchical Faceted Metadata Structures. Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2007: p. 244-251.
68. Anick, P.G. and Tipirneni, S., The paraphrase search assistant: terminological feedback for iterative information seeking, in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 1999, ACM: Berkeley, California, United States. p. 153-159.
69. Ling, X., et al., Mining multi-faceted overviews of arbitrary topics in a text collection, in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008, ACM: Las Vegas, Nevada, USA. p. 497-505.
70. LingPipe Home. Available from: http://www.alias-i.com/lingpipe/.
71. Term Extraction Web Service - YDN. Available from: http://developer.yahoo.com/search/content/V1/termExtraction.html.
72. Taxonomy Warehouse. Available from: http://www.taxonomywarehouse.com/.

73. Roy, S.B., et al., DynaCet: Building Dynamic Faceted Search Systems over Databases, in 2009 IEEE 25th International Conference on Data Engineering(ICDE), Vols 1-3. 2009. p. 1463-1466, 1768.
74. Zhao, B., et al., TEXplorer: keyword-based object search and exploration in multidimensional text databases, in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM). 2011, ACM: Glasgow, Scotland, UK. p. 1709-1718.
75. Zeng, H.-J., et al., Learning to cluster web search results, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004, ACM: Sheffield, UK. p. 210-217.
76. Dou, Z., et al., Finding dimensions for queries, in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM). 2011, ACM: Glasgow, Scotland, UK. p. 1311-1320.
77. Chen, J. and Li, Q., Concept Hierarchy Construction by Combining Spectral Clustering and Subsumption Estimation, in Web Information Systems – WISE 2006, K. Aberer, et al., Editors. 2006, Springer Berlin / Heidelberg. p. 199-209.
78. Atkins, S., Rundell, M., and Sato, H., The contribution of FrameNet to practical lexicography. International Journal of Lexicography, 2003. 16(3): p. 333-357.
79. Sanderson, M. and Croft, B., Deriving concept hierarchies from text. SIGIR'99: Proceedings of 22nd International Conference on Research and Development in Information Retrieval, 1999: p. 206-213, 339.
80. Dakka, W., Ipeirotis, P.G., and Wood, K.R., Automatic construction of multifaceted browsing interfaces, in Proceedings of the 14th ACM Conference on Information and Knowledge Management (CIKM). 2005, ACM: Bremen, Germany. p. 768-775.
81. Xing, D., et al., Deep classifier: automatically categorizing search results into large-scale hierarchies, in Proceedings of the international conference on Web search and web data mining. 2008, ACM: Palo Alto, California, USA. p. 139-148.
82. Krishnapuram, R. and Kummamuru, K., Automatic taxonomy generation: Issues and possibilities, in Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA). 2003. p. 52-63.
83. Holi, M. and Hyvönen, E., Fuzzy View-Based Semantic Search, in The Semantic Web – ASWC 2006, R. Mizoguchi, Z. Shi, and F. Giunchiglia, Editors. 2006, Springer Berlin / Heidelberg. p. 351-365.
84. Roy, S.B., et al., Minimum-effort driven dynamic faceted search in structured databases, in Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM). 2008, ACM: Napa Valley, California, USA. p. 13-22.
85. Yamamoto, T., Nakamura, S., and Tanaka, K., Extracting adjective facets from community Q&A corpus, in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM). 2011, ACM: Glasgow, Scotland, UK. p. 2021-2024.
86. Kleinberg, J.M., Authoritative sources in a hyperlinked environment, in Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms. 1998, Society for Industrial and Applied Mathematics: San Francisco, California, United States. p. 668-677.
87. Uddin, M.N. and Janecek, P., Performance and usability testing of multidimensional taxonomy in web site search and navigation. Performance Measurement and Metrics, 2007. 8(1): p. 18-33.
88. Smith, G., et al., FacetMap: a Scalable Search and Browse Visualization. IEEE Transactions on Visualization and Computer Graphics, 2006. 12(5): p. 797-804.
89. Kekalainen, J., Binary and graded relevance in IR evaluations - Comparison of the effects on ranking of IR systems. Information Processing & Management, 2005. 41(5): p. 1019-1033.

90. Gomadam, K., et al., A Faceted Classification Based Approach to Search and Rank Web APIs, in Proceedings of the 2008 IEEE International Conference on Web Services. 2008, IEEE Computer Society. p. 177-184.
91. Jarvelin, K. and Kekalainen, J., Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 2002. 20(4): p. 422-446.
92. Moffat, A. and Zobel, J., Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems, 2008. 27(1): p. 1-27.
93. Buckley, C. and Voorhees, E.M., Evaluating evaluation measure stability, in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. 2000, ACM: Athens, Greece. p. 33-40.
94. Voorhees, E., The TREC-8 question answering track report, in Proceedings of the 8th Text Retrieval Conference. 1999. p. 77-82.
95. Buckley, C. and Voorhees, E.M., Retrieval evaluation with incomplete information, in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004, ACM: Sheffield, UK. p. 25-32.
96. Macdonald, C., Ounis, I., and Soboroff, I., Overview of the TREC 2009 Web track, in Proceedings of the 18th Text REtrieval Conference (TREC 2009). 2009: Gaithersburg, Maryland, USA.
97. Pound, J., Paparizos, S., and Tsaparas, P., Facet discovery for structured web search: a query-log mining approach, in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011, ACM: Athens, Greece. p. 169-180.
98. Xu, Y. and Mease, D., Evaluating web search using task completion time, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009, ACM: Boston, MA, USA. p. 676-677.
99. Roitman, H., et al., Exploratory search over social-medical data, in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM). 2011, ACM: Glasgow, Scotland, UK. p. 2513-2516.
100. Zhang, J. and Marchionini, G., Evaluation and evolution of a browse and search interface: Relation Browser++, in Proceedings of the 2005 national conference on Digital government research. 2005, Digital Government Society of North America: Atlanta, Georgia. p. 179-188.
101. Grineva, M., et al., Blognoon: exploring a topic in the blogosphere, in Proceedings of the 20th international conference companion on World Wide Web. 2011, ACM: Hyderabad, India. p. 213-216.
102. Hildebrand, M., van Ossenbruggen, J., and Hardman, L., /facet: A Browser for Heterogeneous Semantic Web Repositories, in The Semantic Web - ISWC, I. Cruz, et al., Editors. 2006, Springer Berlin Heidelberg. p. 272-285.
103. SIMILE:Longwell RDF Browser(2003-2005). Available from: http://simile.mit.edu/longwell.
104. Lee, T., et al., Tabulator: Exploring and analyzing linked data on the semantic web, in Procedings of the 3rd International Semantic Web User Interaction Workshop (SWUI). 2006.
105. Huynh, D. and Karger, D., Parallax and Companion: Set-based Browsing for the Data Web, in Proceedings of 18th International World Wide Web Conference. 2009.
106. Dumais, S., et al., Stuff I've seen: a system for personal information retrieval and re-use, in SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. 2003, ACM: Toronto, Canada. p. 72-79.
107. Lee, B., et al., FacetLens: Exposing Trends and Rlationships to Support Sensemaking within Faceted Datasets, in Proceedings of the 27th international conference on Human factors in computing systems. 2009, ACM: Boston, MA, USA. p. 1293-1302.

108. Teregowda, P.B., et al., SeerSuite: developing a scalable and reliable application framework for building digital libraries by crawling the web, in Proceedings of the 2010 USENIX conference on Web application development. 2010, USENIX Association: Boston, MA. p. 14-14.
109. Apache Solr. Available from: http://lucene.apache.org/solr.
110. David Smiley and Pugh, E., Apache Solr 3 Enterprise Search Server. 2011: Packt Publishing. 418.
111. Flamenco. Available from: http://flamenco.berkeley.edu.
112. Nowell, L., Hetzler, E., and Tanasse, T., Change blindness in information visualization: A case study, in Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01) 2001: San Diego, CA, USA. p. 15-22, 171.
113. Stefaner, M., Urban, T., and Seefelder, M., Elastic Lists for Facet Browsing and Resource Analysis in the Enterprise, in Proceedings of the 19th International Conference on Database and Expert Applications Systems. 2008: Turin, Italy. p. 397-401.
114. Simitsis, A., et al., Multidimensional content eXploration. Proc. VLDB Endow., 2008. 1(1): p. 660-671.
115. Chen, H. and Karger, D.R., Less is more: probabilistic models for retrieving fewer relevant documents, in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006, ACM: Seattle, Washington, USA. p. 429-436.
116. Carterette, B. and Chandar, P., Probabilistic models of ranking novel documents for faceted topic retrieval, in Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM). 2009, ACM: Hong Kong, China. p. 1287-1296.