Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web

Jaime Teevan
Microsoft Research
Redmond, WA, USA
teevan@microsoft.com

Susan T. Dumais
Microsoft Research
Redmond, WA, USA
sdumais@microsoft.com

Zachary Gutt
Microsoft Corporation
Redmond, WA, USA
zachg@microsoft.com

ABSTRACT
Faceted search systems help people find what they are looking by allowing them to specify not just keywords related to their information need, but also metadata. While such systems hold great potential and have been successfully used in vertical domains, there are many challenges in extending them to large, heterogeneous collections like the Web, corporate intranets, or federated search engines that access many different data silos. In this position paper we discuss the challenges in greater detail. Those that we have identified stem from the fact that such datasets are 1) ***very large***, making it difficult to assign quality meta-data to every document and to retrieve the full set of results and associated metadata at query time, and 2) ***heterogeneous***, making it difficult to apply the same metadata to every result or every query.

Categories and Subject Descriptors
H.5.4 [Information Systems]: Information Interfaces and Presentation (e.g., HCI) – Hypertext/Hypermedia: User issues.

General Terms: Human Factors, Measurement.

Keywords: Faceted search, filtering, metadata, Web search.

1. INTRODUCTION
The term *facet* means "little face" and is often used to describe one side of a many-sided object, especially a cut gemstone. In the information science literature, the term has been used to refer both to the organization of information (faceted classification), and to interfaces that provide flexible access to that information (faceted search). An important motivation for faceted systems is that any single organizational structure is too limiting. Multiple independent facets provide alternative ways of getting to the same information, thus supporting a wider range of end-user tasks and knowledge. Interfaces to faceted information usually include capabilities for structured browsing (or faceted navigation), and some offer search capabilities as well. In this paper we explore some of the challenges involved in developing faceted search systems for large, unstructured and heterogeneous collections.

The principles of faceted organization are widely applicable. Each facet represents a dimension that can be used to organize the information (e.g., topical category, price, manufacturer, color, etc.). Each facet has a name or label, which can be alphabetic, numeric, categorical, continuous, etc. Facets can be organized hierarchically or as a flat list. Every item in the collection is assigned one or more values on each facet. A probability or confidence can be associated with each value, as often happens when values are assigned automatically, although interfaces that expose this are rare.

Faceted search systems augment full-text search capabilities by providing additional structure to support query refinement or results presentation. Often when people search for information, they prefer to specify as little as necessary in their query to find what they are looking for [1, 2, 8]. Rather than fully specifying their target up front, searchers often prefer to interact with the results to refine their query as necessary. For many search tasks, an initial query is sufficient. When modifications are necessary faceted search provides an easy way for people to further describe what they are looking for. For example, if a person were looking for a $200 red digital camera, instead of typing "$200 red digital camera" into a commerce site's search box, that person may first search for "cameras", and then refine the query by selecting the "digital camera" category, the appropriate price range, and the camera color of their choice. This type of faceted search interaction, which combines full-text search and metadata browsing, has been successfully used in many search verticals, and is commonly seen in e-commerce Web sites, desktop search applications, library databases, etc.

However, there are many challenges to extending the successes of faceted search to large, heterogeneous corpora like the Web, large corporate intranets, or federated search engines that access many different data silos. In this paper, we first summarize some of the lessons learned from previous successful implementations of faceted search in more limited domains, and then discuss some of the challenges faced when scaling up to large, heterogeneous applications.

2. RELATED WORK
Several examples of faceted search systems have been discussed in the research literature, including faceted metadata systems for images [1], movies [5], houses [6], and desktop content [1]. In addition, many Web sites use faceted search to provide access to their content. Examples include: library catalogs (e.g., www2.lib.ncsu.edu/catalog), images (e.g., gettyimages.com), and shopping sites such as BestBuy (bestbuy.com), Home Depot (homedepot.com) and eBay (ebay.com).

Previous research has examined a number of the challenges for developing effective faceted search systems. For example, one issue is how best to represent continuous dimensions. A popular approach is to group continuous facets like "Price" into bins (e.g., $1-$100, $101-$200) that can then be selected. However, bins do not allow users to capture finer distinctions. Shneiderman [6] developed richer interaction techniques that use sliders to highlight ranges of interest and dynamic query techniques to update the display of matching results in real-time.

Another challenge that has been explored is how facets should be combined. Different facets can potentially be specified in any

order and combined to identify a set of items using the full power of Boolean logic. Enabling users to richly express what they are looking for without overwhelming them is an important design goal. In practice, most systems use AND to combine selections from different facets (e.g., red AND $200), and OR to combine selections from the same facet (e.g., (red OR black) AND $200). Hearst [4] provides a nice summary of emerging best practices in user interface design for faceted search, including which facets to show (and how to provide access to others), graphic techniques to display facet labels and matches, and breadcrumb design to indicate the current query terms and facet selections.

In this paper, we discuss additional challenges that may be encountered when applying faceted search to large, heterogeneous corpora. We highlight three issues (generating metadata when it is not explicitly available, identifying which facets to use, and providing quick and accurate metadata profiles), and we look forward to discussing additional issues with workshop attendees.

While there have been attempts to structure the content of the Web using a topic hierarchy like Open Directory (dmoz.org) or the Yahoo! directory in its early days, such systems reflect only a single facet (topic), and the content has not always been tightly integrated with full-text search. Similarly, many search engines provide related searches that allow users to specialize or generalize their requests, but again this exposes only a single dimension (words, which are different in many ways to more traditional facet organizations). Here we focus on the issues related to the tight integration of full-text search and rich faceted navigation.

3. CHALLENGES

The challenges we have identified to applying faceted search to domains like the Web stem from the fact that such datasets are very large and heterogeneous. Because they are very large, it is difficult to assign quality meta-data to every document in the collection and to retrieve the full set of results and their associated metadata at query time. And because they are heterogeneous, it is difficult to apply the same facets to every result or every query. In this section we discuss these issues in greater detail.

3.1 Automatically Generated Metadata

Most domain specific search engines have relatively clean metadata associated with the items in their corpus. For example, commerce search engines tend to be built upon databases with accurate price and brand information. Because other corpora of interest, such as intranets or the Web, do not have pre-assigned metadata, many facets are likely to be assigned algorithmically. This means that some of the metadata may be wrong or have a probabilistic value assigned for it.

When determining how to tune an algorithm that automatically assigns metadata for use in faceted search, it is important to balance the cost of mistakenly assigning a metadata attribute to an information item with the cost of not assigning a piece of metadata to an item when it should be. If selecting a facet yields a lot of unexpected and irrelevant results, users may not find the selection to be worthwhile. On the other hand, if selecting a facet causes many relevant results to be removed from the result set, users may find the risk of missing something valuable to be too high to use the system. Our hypothesis, given the importance of precision in Web search, is that it is better to be accurate than comprehensive, but the right balance surely depends on many factors, including the user's information need, context, and the facet in question.

Rather than making a binary decision that a facet applies to an information item or not, a score can be assigned to indicate the confidence in the assignment. There may be ways to surface this confidence in the assignment of facet labels in a way that makes users comfortable. One possibility is to use a slider that starts with the items that have the highest confidence associated with them and gradually add less certain items. Another place where people appear to have some tolerance for ambiguity is in the ranking of Web search results. Users understand that relevant results are ranked first, less relevant results are ranked later, and that this ranking may or may not be perfectly accurate. Using metadata to support different rankings, rather than to merely filter results, may provide value in some cases. As an example, a person looking to buy a digital camera could search for "digital cameras" and then select "commercial sites" not to filter the results, but rather to rank the results so that those most likely to be commercial are listed first.

Ranking result sets by metadata may prove value, too, in enabling people who are searching very large datasets to better access the long tail. If filtering search results preserves the initial query-based ordering, valuable data that is relevant but ranked relatively low may never be seen. For example, a person who searches for "restaurants" and then filters by "near me" may not want to see the hundreds of restaurants near them ordered by how closely they match the query "restaurants", but rather prefer to see the results ordered by those closest to them.

Another challenge to automatic facet generation is that there are a very large number of different types of facets that one could automatically extract about documents, from simple indications of the presence or absence of a keyword in a document (e.g., "camera"), to much more complex (e.g., synthesizing all of the keywords in the document to determine that it is about "photography"). It is not obvious what level of granularity is appropriate to expose. People may want to interact with fine grain, simple facets that are particularly accurate (e.g., we know for sure if the word "camera" appears in a document), or with concepts that may be less accurate but more expressive. When working with a large number of facets it is also important to identify which facets to surface for a particular query or result set, as we discuss in the next section.

3.2 Identifying which Facets to Surface

Many domain specific search engines, such as ones designed to support commerce searches, recipe searches, or image searches, only need support a relatively narrow range of user tasks. In these cases, it is easy to predict which facets will be the most useful for the searcher. In the case of commerce site, price and brand may be particularly useful, while in recipe search, the ingredients or course may be most useful.

On the other hand, people use more general search engines for a much wider range of complex tasks. On the Web, people conduct research, plan trips, purchase items, and find new jobs using search engines. Similarly, on a corporate intranet people may search for experts, colleague contact information, corporate policies, or valuable research all with the same search engine. When the queries applied to a search system are varied in intent it is unlikely that all facets will apply equally well to all queries. While there may be some commonly useful facets that are always worth displaying, others may need to be selected for display on-the-fly. This raises a number of interesting questions, such as how many facets should be display in a given context, in what order, and, most importantly, how should the most relevant facets be identified.

Facet identification can happen manually or automatically. In the case of manual identification, easy ways must be developed for the user to browse through a large list of potentially irrelevant facets to find the ones they want. One way to winnow this list down may be to eliminate facets that contain no results for the current query. However, as we will discuss later, even this can be a challenge with very large collections of information.

In many cases it may be that people prefer to have the most relevant facets identified for them. The initial query and result set could suggest valuable facets. For example, facets that partition the result set well, facets that are commonly selected for a query, or facets that appear more often than expected may be particularly worth displaying. However the facets that are optimal from a statistical perspective may not correspond to those that the user can best recognize or specify. Additional information may be provided by the user implicitly as they reformulate their query and interact with the result set and the facets. Facets that a particular user has previously found useful may be particularly valuable for that user.

One challenge in dynamically identifying the most appropriate facets for each query and associated result set is that consistency and predictability will be reduced. A more consistent ordering of facets may be useful so that users always know where to find the facets they expect. Or, building on the dynamic menu example, it may be useful to copy split menus [7] and preview a few facets that are particularly likely to be useful while still providing more predictable access to the entire set. Another way to provide some consistency within a task type would be to group facets and trigger the entire group for appropriate queries. For example, a commerce query could trigger a set of facets with price and product information, while a recipe-related query could trigger a set of facets with course and ingredient information.

3.3 Hard to Accurately Preview Facets

Another challenge with supporting faceted search over very large or distributed corpora is that the search engine must be able to quickly compute (or estimate) the facet values for every result that matches a particular query. A search for "tom jones", for example, may returns tens of millions of documents. Most commercial search engines examine only a subset of the possible matches in detail, so it may be difficult to compute the full distribution of facet values for all matching items.

The difficulties in knowing detailed information about the complete result set makes facet identification harder, and potentially more dynamic since the result set available for facet identification changes as the user interacts with it. It can also make previewing facets to give users an idea of what to expect when they select a particular facet challenging. Many faceted search systems preview how many results will be returned if a particular facet is selected. For very large databases, it probably makes sense to abstract this preview to a few discrete buckets (e.g., *one*, *a few*, and *many*), but even a preview intended only to indicate the presence or absence of a result with that facet may be inaccurate. Understanding how to develop algorithms to more accurately predict the distribution of metadata values for a dynamic subset of items (namely those returned for the current search) is a valuable direction for future work.

4. CONCLUSION

Faceted search systems have been used successfully for many vertical applications, including e-commerce, image databases, and library catalogs. In this paper we have discussed some of the challenges that must be faced when considering how to apply ideas from faceted search to support access to large, heterogeneous collections, such as general intranet or Web content. These challenges include how to generate metadata when it is not explicitly available, how to identify which facets to display for a query (and associated result set), and how to provide quick and accurate metadata profiles of the content.

REFERENCES

[1] Cutrell, E., Robbins, D., Dumais, S., and Sarin, R. (2006). Fast, flexible filtering with Phlat. In *Proceedings of CHI '06*, 261-270.

[2] Downey, D., Dumais, S., Liebling, D., and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. To appear in *Proceedings of CIKM'08*.

[3] Dumais, S. (2008). Faceted search. *Encyclopedia of Database Systems.* M. T. Ozsu and L. Liu (Eds.) Springer 2009.

[4] Hearst, M. (2006). Design recommendations for hierarchical faceted search interfaces. In the *SIGIR 2006 Workshop on Faceted Search.*

[5] Koren, J., Zhang, Y, and Liu, X. (2008). Personalized interactive faceted search. In *Proceedings of WWW '08*, 477-486

[6] Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE Software, 11(6),* 70-77.

[7] Sears, A. and Shneiderman, B. (1994). Split menus: Effectively using selection frequency to organize menus. *TOCHI, 1(1),* 27-51.

[8] Teevan, J., Alvarado, C. J., Ackerman, M. A., and Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of CHI `04*, 415-422.

[1] Yee, P., Swearingen, K, Li, K., and Hearst, M. (2003). Faceted metadata for image search and browsing. In *Proceedings of CHI '03,* 401-408.