

Breaking Down the Assumptions of Faceted Search

Vladimir Zelevinsky
Endeca Technologies
101 Main Street
Cambridge, MA 02142
1-617-674-6208
vzelevinsky@endeca.com

ABSTRACT

In this paper, we list several features of faceted search and challenge their implicit assumptions.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *information filtering, query formulation, relevance feedback, retrieval models, search process, selection process*. H.5.2 [Information Interfaces and Presentation]: User Interfaces – *graphical user interfaces (GUI), theory and methods, user-centered design*. I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods – *semantic networks*.

General Terms

Algorithms, Design, Experimentation, Human Factors.

Keywords

Faceted search, user interfaces, refinements, semantic networks, correlation, fuzzy matching.

1. INTRODUCTION

Faceted search has emerged as one of the most effective processes for exploratory search and discovery. It allows the user to locate the records of interest by following, in any order, the process of iterative refinement; it also permits discovering unknown data by exposing the sets of data facets that offer both refinement and data-at-a-glance summarization. When even such a change-resistant organization as the U.S. Government embraces faceted search (<http://www.whitehouse.gov/search>), one can justify considering this process to be the standard way to resolve information retrieval needs.

On one hand, the usage patterns and interface details for faceted search have been ironed out and standardized. Users are starting to use faceted search applications and transferring such acquired knowledge to other applications that utilize the same process [4]. Since faceted search user interfaces exhibit similar look and predictable behavior, the experience of using, for example, HomeDepot.com can be easily replicated at Lowes.com. On the

other hand, like every successful idea, faceted search is headed toward the point in its evolution where it is starting to ossify. The main aim of this paper is, in particular, to list several assumptions of the faceted search experience that are ripe for reconsideration; and, more broadly, to suggest that faceted search, however successful its implementations have been, still contains plenty of unexplored possibilities and can support a plethora of novel applications. The first and third ideas below were successfully prototyped; the second one is currently being investigated.

2. CUSTOM DIMENSIONS

The power of faceted search comes, quite (tauto)logically, from facets: navigable and summarizable properties, tagged onto the records in the system. The problem with such facets is that they have to be created in advance (usually, during data pre-processing), are inflexible (cannot be modified), and might not suit the particular search intent of a given user. While this does apply to numerical properties, the recent advances in analytics allow rapid computation of derived metrics, thus somewhat alleviating the problem (see “Dynamic Facets” section in [1]). With topical (keyword) properties, such as salient natural language terms, however, the issues above fully apply. A text corpus that have been parsed and tagged with typed entities of Person, Organization, and Location type might not suit the needs of the user who is interested in navigating the dimensions of car parts or exploring noteworthy neighborhoods of New York City.

Prior work exists [3, 7, 8] that combines pre-extracted salient terms into topical dimensions; the work in [5] detects particular dimensions that the systems considers useful as leading to potential refinements. We, however, posit the need of a system that is capable of creating such topical dimensions with no pre-processing required whatsoever.

We have created a prototype that allows new dimensions to be created at query-time, combining Endeca (<http://endeca.com/>) structureless database with WordNet semantic network (<http://wordnet.princeton.edu/>). See Figure 1 for the diagram of the user interface and interaction model. In such an interface, the user can enter at query time a seed topic for the automated creation of an additional dimension. This topic term is queried against WordNet, retrieving all its senses. For each sense, we retrieve all related terms by following the meronym, holonym, and hyponym network edges. The results are considered as candidates for our refinements. As the last step, the candidates are checked against the corpus (of course, it is also possible to check the candidates against the current search result / navigation state), by measuring their precision and recall relative to the topic term. The candidates that have sufficiently high f-measures are returned to the user as refinements, along with the counts of matching

documents. When the user clicks on a refinement, the system can perform search for the text of the refinement term – or, if the corpus contains a salient terms dimension and the refinement candidates have been selected from the set of these terms, the action can lead to a traditional dimensional refinement.

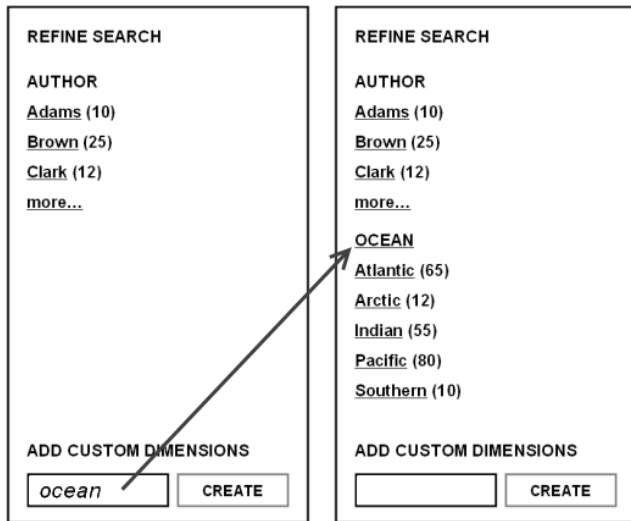


Figure 1. User interface for custom dimensions

The algorithm is fast ($O(N)$, where N is the number of candidate terms) and has the added advantage of providing multiple senses of the topic term, as long as the semantic network contains them.



Figure 2. Two clusters of custom dimensions for “New York”

As an example, see Figure 2. For the user topic “New York”, the system detected two senses (New York as the city vs. New York as the state) and created corresponding refinements. Naturally, the same system can create hierarchical dimensions by recursively submitting each generated refinement as a topic for generation of child refinements.

We currently plan, time permitting, to apply this system to the NYT corpus for the HCIR challenge.

3. DOUBLE DIMENSION SELECTION

The basic tenet of faceted search is a query-response model, where the user reviews the result set as well as the suggested refinements, selects one, and receives the recomputed result set and a new set of refinements from the system. We question the assumption that the user may select only refinement at a time.

There are cases (such as trade-off analysis), where it is desirable to observe the interplay or correlation of several facets before making a selection. Schneiderman [6] proposes a two-dimensional histogram as an indicator of the regions where an intersection of two dimension values does contain corpus data. We propose the user interface where the main element is a two-dimensional scatter plot, which allows the user to see: (1) the distributions of two dimensions of data; (2) their possible correlation; and (3) data density. Additional interface elaborations, such as providing the details for each particular record on mouse-over, are also possible.

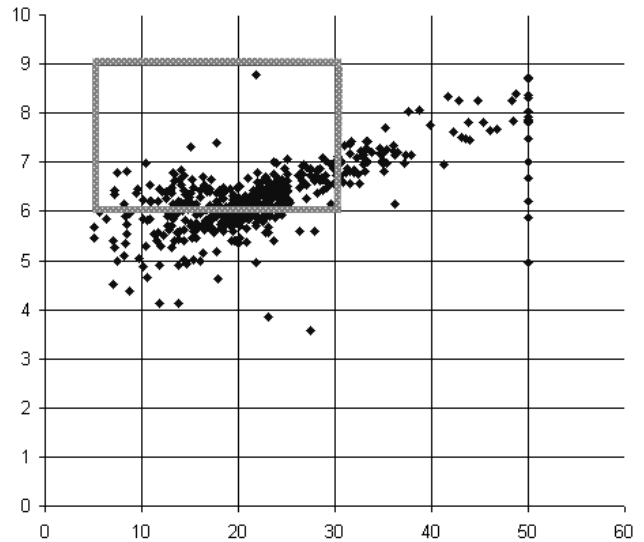


Figure 3. Two-dimensional scatter plot interface

Figure 3 shows the interface, where the data displayed is taken from the well-known 1978 Boston Housing data set (<http://lib.stat.cmu.edu/datasets/boston>): X axis is the median house price (in tens of thousands of dollars) for a town, while Y axis is the average number of rooms. One can easily observe direct correlation between X and Y. The grey rectangle is the user selection that narrows the data to towns with relatively cheap houses with many rooms. Naturally, the application should restrict such selections to regions that do contain data.

Note that the question of which X and Y should be used in construction of the scatter plot interface raises its own set of challenges. We suggest that a system that calculates either Pearson correlation coefficients or K-L divergence for all pairs of dimensions and suggests highly-correlated (or anti-correlated) pairs would be of interest.

This double dimension selection interface happens to share two key properties with other faceted search interfaces: (1) it provides an overview of current result set, while (2) offering ways to refine it. This duality, by the way, might be yet another assumption of faceted search interfaces that is ripe for re-consideration.

4. FUZZY SELECTIONS

Refinements tend to behave as firm restrictions on the result set: the records that do not satisfy them are not included in the refined set. In some cases, this is not the desired behavior. For example, the user might not be familiar with the data and not sure what refinements are relevant for the given search intent. We suggest that in such scenarios it might be helpful to the user to see what records are not inside the selection but are adjacent to it.

There is considerable amount of research on automatic expansion of text queries (see, for example, [2]); we, however, are concerned with the more general case of: (1) automatically expanding the result set for a broadly defined faceted search, where dimensions might be textual, numerical, or discrete, and (2) suggesting to the user those records that are located inside the expanded set but not in the original one.

We (the author along with user interface expert Blade Kotelly and software architect Maia Hansen) created a prototype that ran on top of a restaurant reviews data set. For a given query (which, in the faceted world, is an intersection of refinements), the application relaxed, one at a time, every refinement. Numerical dimensions were relaxed by considering the union of an interval immediately preceding and immediately succeeding the current selection; thus, a selection of “price: \$10-\$20” was substituted with “price: \$5-\$10 OR \$20-\$30”. Discrete dimensions (e.g., “cuisine: Chinese”) were replaced with their inverse (“cuisine: NOT Chinese”). If the relaxed set contained fewer results than the original, we returned them as “Also consider...” suggestions.

The experience of using such a prototype delivered a two-fold reaction. On one hand, the system suggesting a highly-rated Turkish restaurant in the same neighborhood where the user was trying to find a good Greek eatery was akin to experiencing mind-reading. On the other hand, some of the suggestions were perceived as having very little in common with the user query.

The issue of “how one can compare average entrée price with the walking distance to the nearest bus stop” still remains open. As the next step, we are considering applying K-L divergence to detect which properties are the most characteristic of the original

result set (as differing from the complete set of records in the system). Then the expanded set of “penumbra” records can be filtered to select those that share such characteristic properties but might differ on less relevant ones.

5. CONCLUSION

The three assumptions listed above are meant to be neither an exhaustive list – nor even a coherent one, covering as they do pre-processing, refinement selection, and system's response to selected refinements. They are, however, intended to indicate several ways of breaking down standard assumptions and conventions of the faceted search interface. If this paper succeeds in encouraging UX researchers to “think outside the refinement box” and look for under-explored possibilities of faceted search, it will have fulfilled its intended function.

6. REFERENCES

- [1] O. Ben-Yitzhak et al. Beyond Basic Faceted Search. WSDM 2008. DOI: <http://doi.acm.org/10.1145/1341531.1341539>
- [2] C. Carpineto, R. de Mori, G. Romano, B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, volume 19, 1-27, January 2001. DOI: <http://doi.acm.org/10.1145/366836.366860>
- [3] W. Dakka, R. Dayal, P. Ipeirotis. Automatic Discovery of Useful Facet Terms. *Proceedings of the ACM SIGIR '06 Workshop on Faceted Search*, 2006.
- [4] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, K.-P. Yee. Finding the flow in web site search. *Communications of the ACM*, 45:42-49, September 2002. DOI: <http://doi.acm.org/10.1145/567498.567525>
- [5] C. Li, N. Yan, S. B. Roy, L. Lisham, G. Das. Facetedpedia: Dynamic Generation of Query Dependent Faceted Interfaces for Wikipedia. *International World Wide Web Conference*, Raleigh, North Carolina, USA, 2010. DOI: <http://doi.acm.org/10.1145/1772690.1772757>
- [6] B. Schneiderman, Dynamic Queries for Visual Information Seeking. *IEEE Software*, 11, 70-77, June 1994. DOI: <http://doi.acm.org/10.1109/52.329404>
- [7] E. Stoica, M. Hearst. Demonstration: Using WordNet to Build Hierarchical Facet Categories. *ACM SIGIR Workshop on Faceted Search*, August, 2006.
- [8] K. Yang, E. Jacob, A. Loehrlein, S. Lee, N. Yu. Organizing the Web: semi-automatic construction of a faceted scheme. *IADIS International Conference WWW/Internet*, Madrid, Spain, 2004.