Custom Dimensions for Text Corpus Navigation

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

ABSTRACT

We report on our Custom Dimension search application, built for HCIR Challenge on the basis of the New York Times annotated corpus, Endeca structureless database, and WordNet semantic network.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, query formulation, retrieval models, search process, selection process*. I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *semantic networks*.

General Terms

Algorithms, Design, Experimentation.

Keywords

Faceted search, user interfaces, refinements, semantic networks.

1. INTRODUCTION

The power of faceted search comes from facets: navigable and summarizable properties, tagged onto the records in the system. The problems with facets are: 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, the issues above present real problems. A text corpus that has been parsed and tagged with typed entities of Person, Organization, and Location type might not suit the needs of a user who is interested in navigating the dimensions of car parts or exploring noteworthy neighborhoods of New York City.

Prior work exists [2, 4, 5] that combines pre-extracted salient terms into topical dimensions; the work in [3] 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.

2. PROTOTYPE

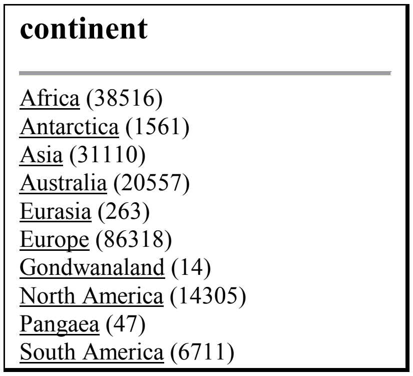
We have created a prototype application that allows new dimensions to be created at query-time, combining Endeca (http://endeca.com/) structureless database with WordNet semantic network (http://wordnet.princeton.edu/), and applied the resulting application to the New York Times annotated corpus (http://corpus.nytimes.com/). In our 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 (e.g., 1: "New York" as a city, and 2: "New York" as a state). For each sense, the application retrieves 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. After experimenting with several variations on the f-measure, we ended up simply using the frequency of the candidate term in the entire corpus as the sole relevance criterion. When the user clicks on a refinement, the system performs search for the text of the refinement term on the body of the articles.

The algorithm is fast and has the added advantage of providing multiple senses of the topic term, as long as the semantic network contains them. In our application, we intentionally disabled all other refinements in order to showcase the power of custom dimensions.

We would be glad to present a live demo at HCIR if invited.

3. RESULTS

The custom dimensions interface shares two key properties with other faceted search interfaces: (1) it provides an overview of current result set, while (2) offering ways to refine it. To our surprise, the interface is useful in a third way: it helps the users to reconsider their initial assumptions, thus allowing broadening of the search intent / task. As an example, see Figure 1, where we display custom dimensions for the term "continent". One would expect to see "Africa" and "Asia"; one is less likely to expect seeing "Gondwanaland" and "Pangea" on the same list. In our (informal) usability study, we found this property of the interface repeatedly affecting the users' information retrieval process.

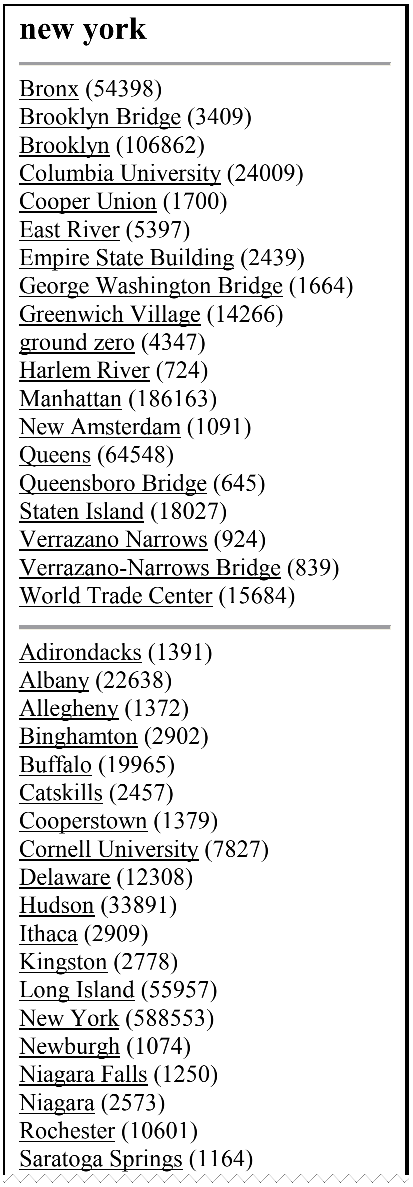


continent

Africa (38516)
Antarctica (1561)
Asia (31110)
Australia (20557)
Eurasia (263)
Europe (86318)
Gondwanaland (14)
North America (14305)
Pangaea (47)
South America (6711)

Figure 1. Custom dimensions for "continent"

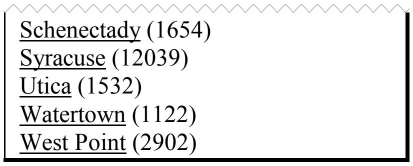
Figure 2 displays the dimensions for the user topic "New York". Here, the system detected two senses (New York as the city vs. New York as the state) and created corresponding refinements.



new york

Bronx (54398)
Brooklyn Bridge (3409)
Brooklyn (106862)
Columbia University (24009)
Cooper Union (1700)
East River (5397)
Empire State Building (2439)
George Washington Bridge (1664)
Greenwich Village (14266)
ground zero (4347)
Harlem River (724)
Manhattan (186163)
New Amsterdam (1091)
Queens (64548)
Queensboro Bridge (645)
Staten Island (18027)
Verrazano Narrows (924)
Verrazano-Narrows Bridge (839)
World Trade Center (15684)

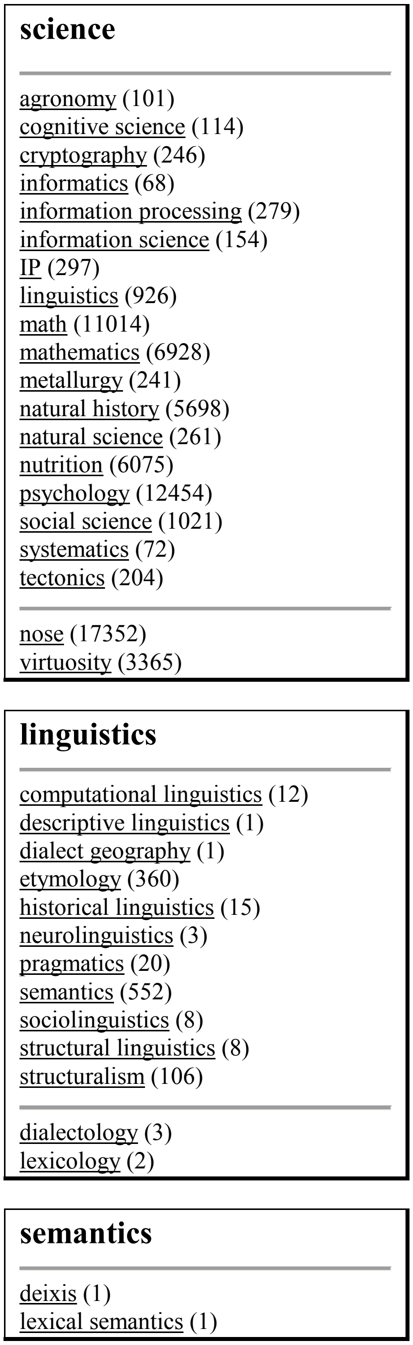
Adirondacks (1391)
Albany (22638)
Allegheny (1372)
Binghamton (2902)
Buffalo (19965)
Catskills (2457)
Cooperstown (1379)
Cornell University (7827)
Delaware (12308)
Hudson (33891)
Ithaca (2909)
Kingston (2778)
Long Island (55957)
New York (588553)
Newburgh (1074)
Niagara Falls (1250)
Niagara (2573)
Rochester (10601)
Saratoga Springs (1164)



Schenectady (1654)
Syracuse (12039)
Utica (1532)
Watertown (1122)
West Point (2902)

Figure 2. Custom dimensions for "New York"

Figure 3 displays the dimensions for "science"; its child, "linguistics"; and its child, "semantics", illustrating the possibility of creating custom hierarchies.



science

agronomy (101)
cognitive science (114)
cryptography (246)
informatics (68)
information processing (279)
information science (154)
IP (297)
linguistics (926)
math (11014)
mathematics (6928)
metallurgy (241)
natural history (5698)
natural science (261)
nutrition (6075)
psychology (12454)
social science (1021)
systematics (72)
tectonics (204)

nose (17352)
virtuosity (3365)

linguistics

computational linguistics (12)
descriptive linguistics (1)
dialect geography (1)
etymology (360)
historical linguistics (15)
neurolinguistics (3)
pragmatics (20)
semantics (552)
sociolinguistics (8)
structural linguistics (8)
structuralism (106)

dialectology (3)
lexicology (2)

semantics

deixis (1)
lexical semantics (1)

Figure 3. Custom dimensions for "science" and its children

4. CONCLUSIONS

The main issue with using the custom dimensions application is the polysemic nature of language: when performing text search for a term, it is not guaranteed the system will return only the documents where this term is used in the sense that corresponds to the given dimension topic; for example, selecting “Queens” from the “New York” dimension (Figure 2) will also return documents that refer to monarchy. Restricting matches to proper nouns (possibly even pre-extracted with a Location entity extractor) will solve this particular issue, but will not help with the case of “Syracuse” being not only a city in the state of New York, but also a town in Sicily. A possible solution is a reduced-recall, increased-precision replacement: instead of searching for the refinement term, the application can perform the search for the term as well as the text of the topic seed.

The possibility of limiting refinement to pre-extracted salient terms likewise remains a promising venue of investigation.

The author would like to thank The New York Times for providing the corpus, and Adam Ferrari for his assistance with data wrangling.

5. REFERENCES

[1] O. Ben-Yitzhak et al. Beyond Basic Faceted Search. WSDM 2008. DOI: http://doi.acm.org/10.1145/1341531.1341539

[2] W. Dakka, R. Dayal, P. Ipeirotis. Automatic Discovery of Useful Facet Terms. *Proceedings of the ACM SIGIR ‘06 Workshop on Faceted Search*, 2006.

[3] 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

[4] E. Stoica, M. Hearst. Demonstration: Using WordNet to Build Hierarchical Facet Categories. ACM SIGIR Workshop on Faceted Search, August, 2006.

[5] 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.