NLP Support for Faceted Navigation in Scholarly Collections

Marti A. Hearst
School of Information, UC Berkeley
102 South Hall, Berkeley, CA 94720
hearst@ischool.berkeley.edu

Emilia Stoica
Ask.com
555 12th Street, Oakland, CA 94607
emilia.stoica@ask.com

Abstract

Hierarchical faceted metadata is a proven and popular approach to organizing information for navigation of information collections. More recently, digital libraries have begun to adopt faceted navigation for collections of scholarly holdings. A key impediment to further adoption is the need for the creation of subject-oriented faceted metadata. The Castanet algorithm was developed for the purpose of (semi) automated creation of such structures. This paper describes the application of Castanet to journal title content, and presents an evaluation suggesting its efficacy. This is followed by a discussion of areas for future work.

1 Introduction

Faceted navigation for searching and browsing "vertical" content collections has become the standard interface paradigm for e-commerce shopping web sites. Faceted navigation, when properly designed, has been shown to be understood by users and preferred over other organizations (Hearst et al., 2002; Yee et al., 2003; English et al., 2001). Although text clustering is an easily automated technique, numerous studies have found that the results of clustering are difficult for lay people to understand (Kleiboemer et al., 1996; Russell et al., 2006; Hornbæk and Frøkjær, 1999) and that the coherent and predictable structure of categorical metadata is superior from a usability prespective (Rodden et al., 2001; Pratt et al., 1999; Hearst, 2006a).

An interface using hierarchical faceted navigation simultaneously shows previews of where to go next and how to return to previous states in the exploration, while seamlessly integrating free text search within the category structure. Faceted metadata provides organizing context for results and for subsequent queries, which can act as important scaffolding for exploration and discovery. The mental work of searching an information collection is reduced by promoting recognition over recall and suggesting logical but perhaps unexpected alternatives, while at the same time avoiding empty results sets.

Recently, faceted navigation has emerged as the dominant method for new interfaces for navigating digital library collections. The NCSU library catalog was an early adopter among university libraries, using the Endeca product as its backend (Antelman et al., 2006). A usability study with 10 undergraduates comparing this system to the old library catalog interface found a 48% improvement in task completion time, although the study did not account for the effects of facets vs. the effects of fuller coverage in the keyword search.

Additionally, a consortium of university libraries (the OCLC) is now using the WorldCat shared catalog and interface, which features a faceted navigation component (see Figures 1 and 2). And another popular interface solution is provided by AquaBrowser, in this case, shown on the University of Chicago website (see Figure 3). A recent study on this site found significant benefits attributable to the faceted navigation facility (Olson, 2007). And finally, the online citation system DBLP has not one but two different faceted interfaces, as does the ACM Digital Library.

These interfaces do a good job of allowing users to filter by bibliographic attributes such as media, date, and library. However, in most cases the subject metadata still is not as rich as it should be to fully facilitate information browsing and discovery in these systems. In fact, there are a number of open problems with the use of faceted navigation for scholarly work. Some of these have to do with how best to present faceted navigation in the interface (Hearst, 2006b), but others are more relevant

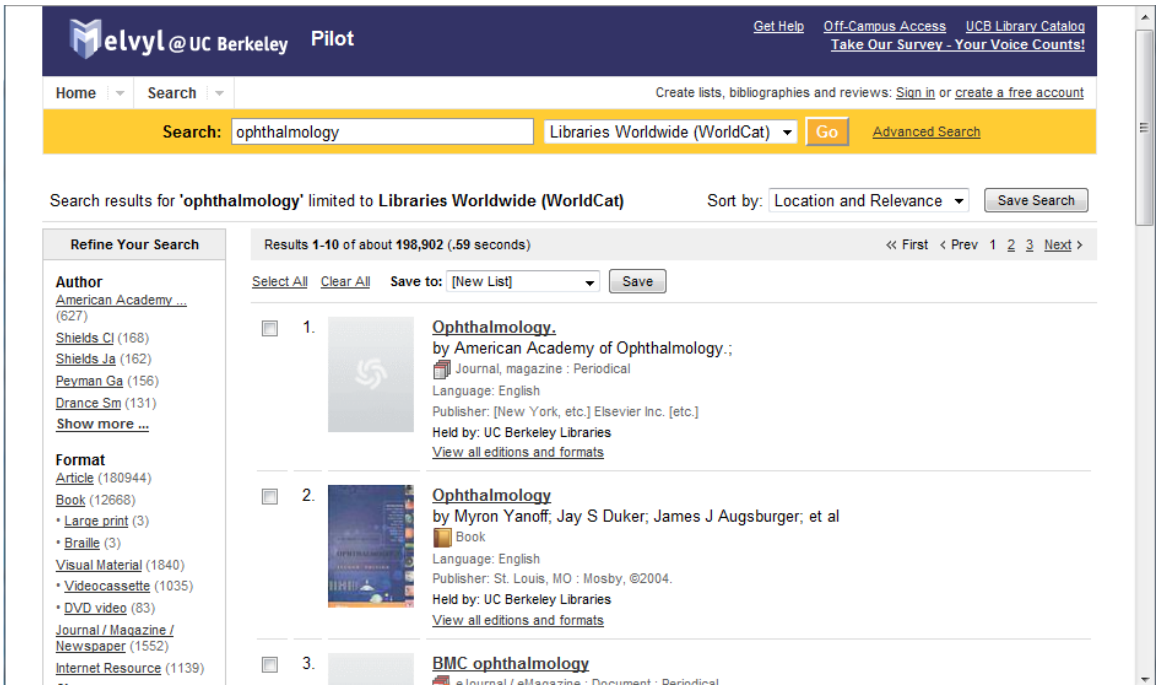

Figure 1: Worldcat consortium digital library interface using faceted navigation. The instance shown is the University of California version, from http://berkeley.worldcat.org .

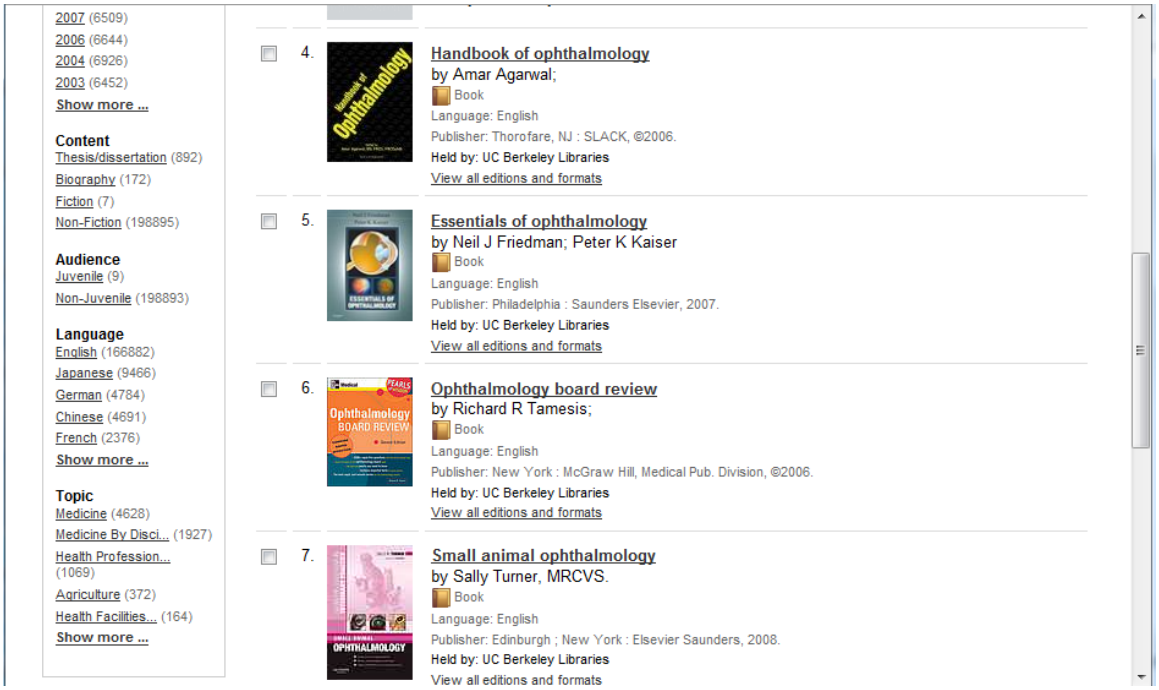

Figure 2: Digital library interface with faceted navigation, continued, from http://berkeley.worldcat.org .

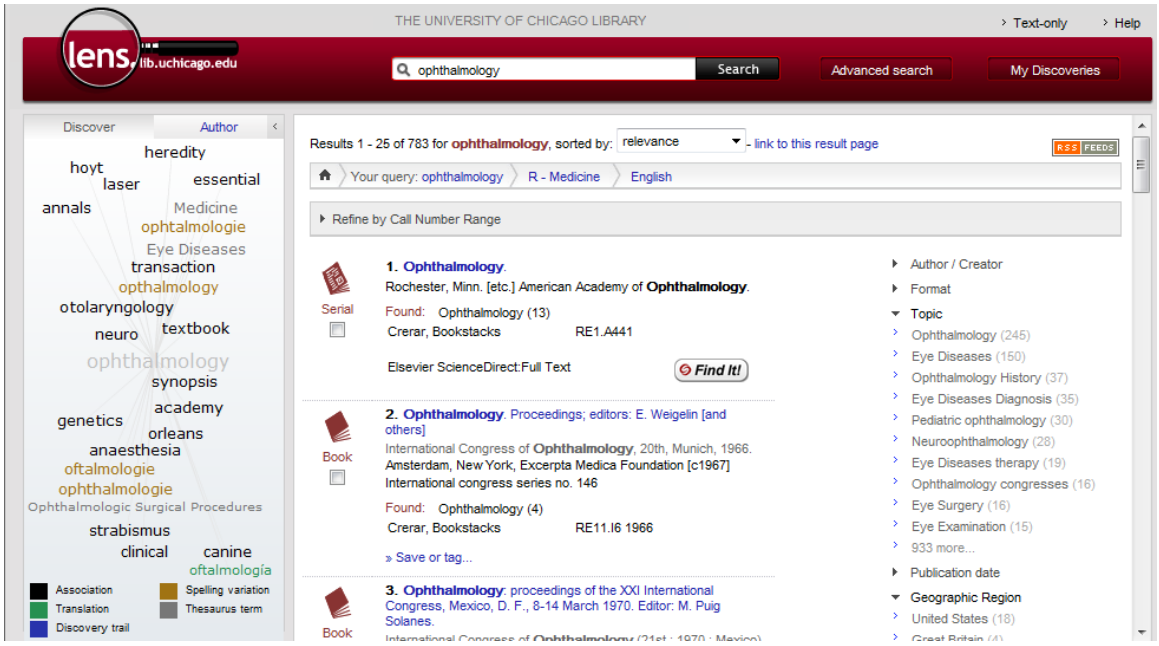

Figure 3: University of Chicago digital library interface using faceted navigation, using an interface from AquaBrowser.

to NLP, including:

- How to automatically or semi-automatically create rich subject-oriented faceted metadata for scholarly text?
- How to automatically assign information items to faceted category labels?

This paper describes the results of applying Castanet, a semi-automated approach to creating faceted metadata, to a scholarly collection. (In past work it has been shown to work well on a different kind of text (Stoica et al., 2007; Stoica and Hearst, 2004).) It then discusses some open problems in building navigation structures for scholarly digital libraries.

2 Creating Faceted Metadata

This section first defines faceted metadata, and then describes the CastaNet algorithm. More details about the algorithm can be found in a prior publication (Stoica et al., 2007).

Rather than one large category hierarchy, faceted metadata consists of a set of categories (flat or hierarchical), each of which corresponds to a different facet (dimension or feature type) relevant to the collection to be navigated. After the facets are designed, each item in the collection is assigned any number of labels from the facets.

Faceted metadata is intermediate in complexity between flat categories and full knowledge representation. The idea is to develop a set of "orthogonal" categories that characterize the information space in a meaningful way, using terminology that is useful for browsing the contents of a domain. Each facet is a different topic, subject, attribute, or feature, and some facets have hierarchical "is-a" structure. For instance, the facets of a biomedical collection should cover disease, anatomy, drugs, symptoms, side-effects, properties of experimental subjects, and so on. Each biomedical article can then be assigned any number of category labels from any number of facets. An article on the effects of tamoxifen on ovarian cancer when tested on mice could then be navigated to by first starting with cancer, then selecting drug tamoxifen, and then body part ovary, or first with tamoxifen, then navigating to ovary, and further refining by disease type. This ability to "mix and match" both for describing the articles and for navigating the category structure is key.

The term "faceted classification" was deliberately chosen in the Flamenco project to echo the old library science term of that name (Hearst, 2000), but with a rejection of the strict terms required for construction of controlled vocabulary, which mandates exhaustive, mutually exclusive category composition. Rather, the faceted naviga-

tion approach for design of search interfaces calls for category systems that are expressed at a meaningful level of description, use approachable language (unless designed for specialists), are consistent in terms of specificity at each level, avoiding becoming too broad or too deep.

The most difficult part of the design is determining whether or not compound concepts should be created. For instance, when evaluating tags for a digital library like librarything, should terms like "african history" and "british literature" be separated into two facets, one containing major writing types (history, literature), and another nationalities (african, british), or should the modifying structure be retained, as there are many kinds of history and many kinds of literature? Most likely, the answer should depend on the makeup of the collection and the usage that the users are expected to want to make of it.

The next subsections briefly describe related work in automated creation of structure from text, the Castanet algorithm and its output on journal article title text, and the results of a usability study on this output.

2.1 Related Work

One way to create faceted metadata is to start with existing vocabularies, and in fact work has been done on this area. The Library of Congress Subject headings are shown in the U Chicago catalog, despite a statement by Antelman et al. (2006) about the "unsuitability of Library of Congress Subject Headings (LCSH) as an entry vocabulary." There has also been work on converting LCSH into faceted metadata (Anderson and Hofmann, 2006). Work on the Flamenco project converted the Art and Architecture thesaurus to a faceted category system manually (Hearst et al., 2002). However, automated techniques are desirable.

Other methods that are influential but claimed to make a meaningful category structure, but not necessarily a faceted one, include the LDA (Latent Dirichlet Allocation) method (Blei et al., 2003), which uses a generative probabilistic model of discrete data to create a model of documents' topics. It attempts to analyze a text corpus and extract the topics that combine to form the documents. The output of the algorithm was originally evaluated in terms of perplexity reduction but not in terms of understandability of the topics produced.

Sanderson and Croft (1999) propose a method called Subsumption for building a hierarchy for a set of documents retrieved for a query. For two terms *x* and *y*, *x* is said to subsume *y* if the following conditions hold: $P(x|y) \geq 0.8, P(y|x) < 1$. To evaluate the algorithm the authors asked 8 participants to look at parent-child pairs and state whether or not they were "interesting." Participants found 67% to be interesting as compared to 51% for randomly chosen pairs of words. Of those interesting pairs, 72% were found to display a "type-of" relationship.

Another class of solutions make use of existing lexical hierarchies to build category hierarchies, as we do in this paper. For example, Navigli and Velardi (2003) use WordNet (Fellbaum, 1998) to build a complex ontology consisting of a wide range of relation types (demonstrated on a travel agent domain), as opposed to a set of human-readable hierarchical facets. Mihalcea and Moldovan (2001) describe a sophisticated method for simplifying WordNet in general, rather than tailoring it to a specific collection.

Zelevinsky et al. (2008) used an approach of looking at keywords assigned by authors of ACM publications to documents, computing which terms had high importance within those documents, and then using the highest scoring among those documents to assign new keywords (referred to in the paper as tags) to the documents. The tags were shown as query term refinements in a digital library interface.

Only limited related work has attempted to make faceted category hierarchies explicitly. Dakka et al. (Dakka and Ipeirotis, 2008; Dakka et al., 2005) is one of these. Their approach is a combination of Subsumption and Castanet; they use lexical resources like WordNet and Wikipedia to find structure among words, but also use them to determine which words in a collection are most useful to include in a faceted system. The facet hierarchy is made via Subsumption. The evaluation of their most recent work on news text finds strong results for assessments made by judges of precision and recall. Furthermore, when facets were shown in a search interface to five users, the keyword usage dropped in favor of clicking on categories, as task completion time was reduced while satisfaction remained unchanged. No examples of facet categories produced by the algorithm are shown, and the role of hierarchy is not clear, but the approach appears especially promising for de-

termining which words of long documents to include in building facet systems.

2.2 Castanet Applied to Journal Titles

The main idea behind the Castanet algorithm is to carve out a structure from the hypernym ("is-a") relations within the WordNet (Fellbaum, 1998) lexical database (Stoica et al., 2007; Stoica and Hearst, 2004). The Castanet algorithm assumes that there is text associated with each item in the collection, or at least with a representative subset of the items. The textual descriptions are used *both* to build the facet hierarchies and to assign items (documents, images, citations, etc.) to the facets, and the text can be fragmented.

The algorithm has five major steps which are briefly outlined here. For details, see (2007).

1. Select target terms from textual descriptions of information items.
2. Build the Core Tree:
    - For each term, if the term is unambiguous, add its synset's IS-A path to the Core Tree.
    - Increment the counts for each node in the synset's path with the number of documents in which the target term appears.
3. Augment the Core Tree with the remaining terms' paths:
    - For each candidate IS-A path for the ambiguous term, choose the path for which there is the most document representation in the Core Tree.
4. Compress the augmented tree.
5. Remove top-level categories, yielding a set of facet hierarchies.

In addition to augmenting the nodes in the tree, adding in a new term increases a count associated with each node on its path; this count corresponds to how many documents the term occurs in. Thus the more common a term, the more weight it places on the path it falls within. The Core Tree acts as the "backbone" for the final category structure. It is built by using paths derived from unambiguous terms, with the goal of biasing the final structure towards the appropriate senses of words. Currently a word can appear in only one sense in the final structure; allowing multiple senses is an area of research.

Figures 4 and 5 show the output of the Castanet algorithm when applied to the titles of journals from the bioscience literature. Note that even the highly ambiguous common anatomy words are successfully grouped using this algorithm, presumably because of the requirement that each word occur in only one location in the ontology and because the anatomy part of the ontology is strongly favored during the part of the process in which the core tree is built with unambiguous terms. (Although some versions of Castanet use an advanced version of WordNet Domains (Magnini, 2000), they were not used in the construction of this category set.)

As reported earlier (Stoica et al., 2007), an evaluation of this algorithm was conducted by asking information architects with expertise in the domain over which the algorithm was run to state whether or not they would like to use the output of the algorithm to build a website. The output of Castanet was compared to Subsumption (Sanderson and Croft, 1999) and to LDA (Blei et al., 2003).

As reported earlier, on a recipes collection, all 34 information architects overwhelming preferred Castanet. They were asked to respond to how likely they would be to use the output, on a scale of: definitely no, no, yes, definitely yes. For Castanet, 85% of the evaluators said yes or definitely yes for intent to use. Subsumption received 38% answering yes or definitely yes, and LDA was rejected by all participants.

The study was also conducted using a biological journal titles collection. 3275 titles were used (although a significant number are not in English and so many are missed by the algorithm). The 15 participants who evaluated the Biomedical titles collection were required to be frequent users of PubMed (the online library for biomedicine), but were not required to be information architects, as it was difficult to finding information architects with biological expertise. These participants were biologists, doctors, medical students and medical librarians.

7 participants saw both LDA and Castanet, and 8 participants saw both Subsumption and Castanet (a pilot test found that participants who saw both Subsumption and LDA became very frustrated with the tasks, so the two options were compared pairwise to Castanet for subsequent trials). For Castanet, 11 out of 15 participants (73%) an-

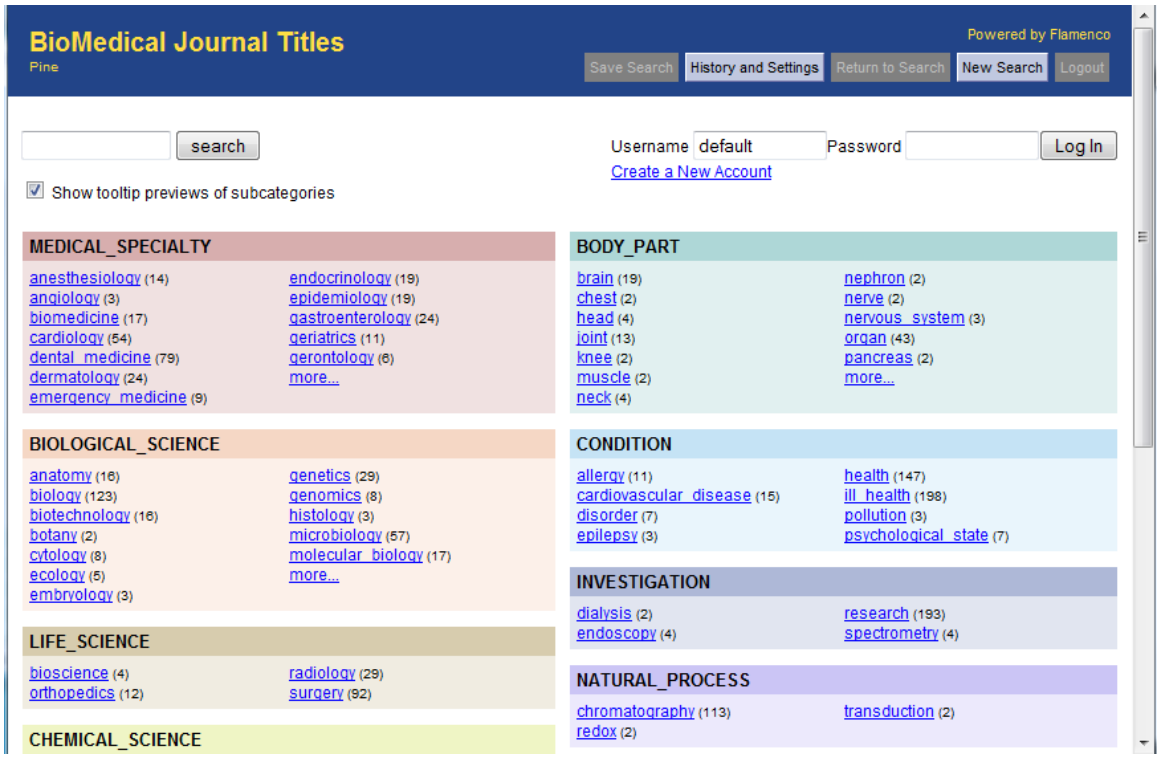

Figure 4: Castanet output on journal title text.

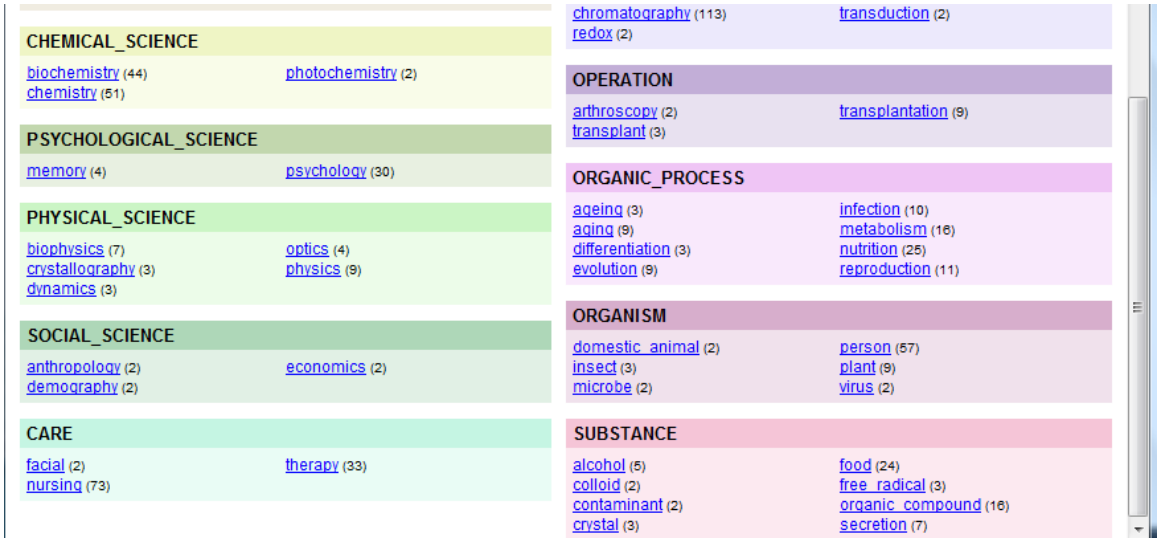

Figure 5: Castanet output on journal title text, continued.

Figure 6: LDA output on journal title text.

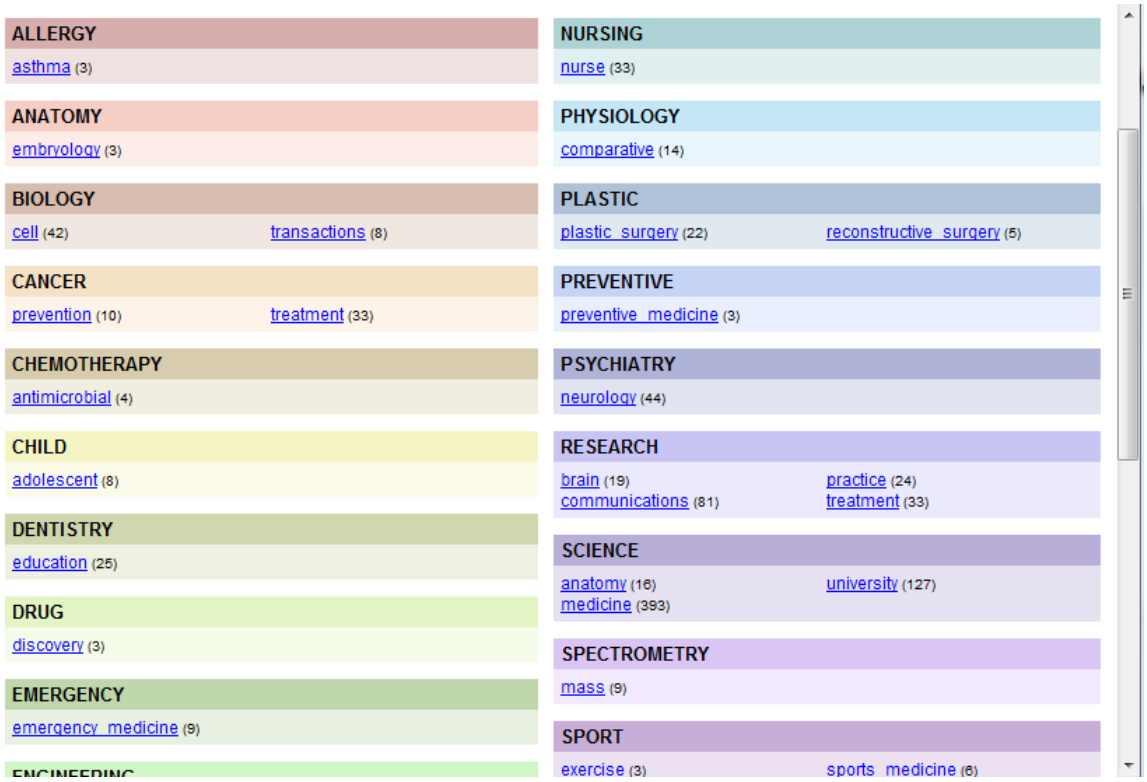

Figure 7: Subsumption output on journal title text.

swered yes or definitely yes to a desire to use its output. 1 out of 7 participants answered yes to a desire to use LDA, and 1 out of 8 answered yes to Subsumption. LDA received 4 "definitely no" responses, whereas Subsumption received only one of these, and no one said definitely no to Castanet.

2.3 Open Problems

Although quite useful "out of the box," the Castanet algorithm could benefit by several improvements and additions:

1. The processing of the terms should recognize spelling variations (such as aging vs. ageing) and morphological variations. Verbs and adjectives are often quite important for a collection and should be included, but with caution.
2. In a related point, the system should have a way of suggesting synonyms to annotate a given node, as opposed to listing closely related words as children or siblings of one another.
3. Some terms should be allowed to occur with more than one sense if this is required by the dataset. For example, the term *brain* is annotated with two domains, *Anatomy* and *Psychology*, which are both relevant domains for a biomedical journal collection.
4. Words that appear in noun compounds and phrases that are not in WordNet should receive special processing.
5. Currently if a term is in a document it is assumed to use the sense assigned in the facet hierarchies; this is often incorrect, and so terms should be disambiguated within the text before automatic category assignment is done.
6. WordNet is not exhaustive and some mechanism is needed to improve coverage for unknown terms.
7. Castanet seems to work better when applied to short pieces of text (e.g., journal titles vs. full text); to remedy this, better methods are needed to select the target terms.
8. A method for dynamically adding facets and adding terms to facets should be developed, especially a method for allowing user tags to be incorporated into the exising facet hierarchies.

Recent work by Dakka et al. (Dakka and Ipeirotis, 2008) can help with point 7, and some recent work by Koren et al. (Koren et al., 2008) seems promising for 8.

Robust evaluation methods are also needed; making use of log information about which facets are heavily used can help inform decisions about which facets work well and which need modification or additions.

Acknowledgements: Megan Richardson provided valuable contributions in her work on the study reported on here. Emilia Stoica did this work while a postdoctoral researcher at UC Berkeley.

References

J.D. Anderson and M.A. Hofmann. 2006. A fully faceted syntax for Library of Congress subject headings. *Cataloging & Classification Quarterly*, 43(1):7–38.

K. Antelman, E. Lynema, and A.K. Pace. 2006. Toward a twenty-first century library catalog. *Information technology and libraries*, 25(3):128–138.

David Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

W. Dakka and P.G. Ipeirotis. 2008. Automatic extraction of useful facet hierarchies from text databases. In *IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008*, pages 466–475.

W. Dakka, P.G. Ipeirotis, and K.R. Wood. 2005. Automatic construction of multifaceted browsing interfaces. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 768–775. ACM New York, NY, USA.

J. English, M.A. Hearst, R. Sinha, K. Swearingen, and K.-P. Yee. 2001. Examining the usability of web site search. Unpublished Manuscript, http://flamenco.berkley.edu/papers/epicurious-study.pdf.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

M.A. Hearst, J. English, R. Sinha, K. Swearingen, and K.-P. Yee. 2002. Finding the flow in web site search. *Communications of the ACM*, 45(9), September.

M.A. Hearst. 2000. Next Generation Web Search: Setting Our Sites. *IEEE Data Engineering Bulletin*, 23(3):38–48.

M.A. Hearst. 2006a. Clustering Versus Faceted Categories For Information Exploration. *Communcations Of The Acm*, 49(4):59–61.

M.A. Hearst. 2006b. Design recommendations for hierarchical faceted search interfaces. In *SIGIR'06 Workshop On Faceted Search*, Seattle, Wa, August.

K. Hornbæk and E. Frøkjær. 1999. Do Thematic Maps Improve Information Retrieval. *Human-Computer Interaction (INTERACT'99)*, pages 179–186.

A.J. Kleiboemer, M.B. Lazear, and J.O. Pedersen. 1996. Tailoring a retrieval system for naive users. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR '96)*, Las Vegas, NV.

J. Koren, Y. Zhang, and X. Liu. 2008. Personalized interactive faceted search. *WWW '08: Proceeding of the 17th international conference on World Wide Web*.

Bernardo Magnini. 2000. Integrating subject field codes into WordNet. In *Proc. of LREC 2000*, Athens, Greece.

Rada Mihalcea and Dan I. Moldovan. 2001. Ez.wordnet: Principles for automatic generation of a coarse grained wordnet. In *Proc. of FLAIRS Conference 2001*, May.

Roberto Navigli, Paola Velardi, and Aldo Gangemi. 2003. Ontology learning and its application to automated terminology translation. *Intelligent Systems*, 18(1):22–31.

T.A. Olson. 2007. Utility of a faceted catalog for scholarly research. *Library Hi Tech*, 25(4):550–561.

W. Pratt, M.A. Hearst, and L. Fagan. 1999. A knowledge-based approach to organizing retrieved documents. In *Proceedings of 16th Annual Conference on Artificial Intelligence(AAAI 99)*, Orlando, FL.

K. Rodden, W. Basalaj, D. Sinclair, and K. R. Wood. 2001. Does organisation by similarity assist image browsing? In *Proceeedings of ACM CHI 2001*, pages 190–197.

D.M. Russell, M. Slaney, Y. Qu, and M. Houston. 2006. Being literate with large document collections: Observational studies and cost structure tradeoffs. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*.

Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of SIGIR 1999*.

E. Stoica and M. Hearst. 2004. Nearly-automated metadata hierarchy creation. In *Companion Proceedings of HLT-NAACL'04*, pages 117–120.

E. Stoica, M.A. Hearst, and M. Richardson. 2007. Automating Creation of Hierarchical Faceted Metadata Structures. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 244–251.

K.-P. Yee, K. Swearingen, K. Li, and M.A. Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of ACM CHI 2003*, pages 401–408. ACM New York, NY, USA.

V. Zelevinsky, J. Wang, and D. Tunkelang. 2008. Supporting Exploratory Search for the ACM Digital Library. In *Workshop on Human-Computer Interaction and Information Retrieval (HCIR'08)*.