Flexible Facets Generation for Faceted Search

Gan Keng Hoon and Teh Chek Wei

School of Computer Sciences
Universiti Sains Malaysia, 11800 Penang, Malaysia
khgan@usm.my

Abstract. Most existing faceted search systems enable filtering of information via facet and values which are derived from structured form. However, in many domains, there are interesting aspects where information can be further refined based on the contents of the domain. This contents are unstructured hence poses challenge to extract the value dynamically from them. This paper showcases a faceted search system that enables generation of facets and values dynamically from unstructured text to further improve the results filtering of retrieval process. The system is demonstrated in the context of bibliographical search.

Keywords: Faceted Search, Bibliography, Search Engine

1 Introduction

Faceted search is a way of organizing information into facet and value pair so that they can be used by users to narrow the search results based on the criteria (i.e. facet and values) they are interested in [1], [2]. For example, facets like year, price, age etc. are used in filtering of product instances; facets like pros and cons are used in filtering review aspects; facets like venue, type, co-authors are used in filtering publications/articles and so forth. In general, facets resemble domain specific categories; whereas facet's values are description related to instances or document unit classified under these categories. For example, a facet, "year", can have values like "2015", "2008" etc. which are related to particular instances like publication. A venue "facet" can have values like "CIKM", "SIGIR" etc. An instance can have more than one facets for different combinations of filtering.

Normally, facets are linked to structured data source, e.g. when we look up a facet, "year"; this facet is linked to the year data [3], [4]. These filters are often based on structured information (e.g. year, venue, price, size etc.) like field name and values in a database table, element name and content in an XML file. However, as these filters are based on only structured information, this limits the capability for selecting instances based on unstructured text related to the instances, which can be very important as well. For example, in the domain of bibliographical search, there are other aspects whether information can be discovered.

In order to enable better search filtering, there are many domain oriented facets which can be adopted. For example, in the search for an expert's bibliographical information, there are specific facets like algorithm, framework, model, application etc. which may be useful to refining the search results. Nevertheless, the generation of facets values can be challenging as these values are normally not specified as part of

the structured/descriptive information of a bibliographical text. However, the positive side is that these values are contained in the text, e.g. “Hidden Markov Model” contains information about the model used (see Fig. 1). Some related works have attempted to improve the filtering through usage of concept/textual features [5], [6] or semantic structure like ontology/taxonomy [7], [8] etc. However, these works are able to capture the concepts in general without linking to certain aspects of domain needs. Hence, in this paper, we focus on improving the existing faceted search by allowing flexible defining of domain specific facets followed by extraction of corresponding values automatically. The extracted facets and values are applied in the context of bibliographical search.

Fig. 1. Structured (grey box) vs. unstructured (black box) facets and their values.

2 Architecture

In this section, we provide a brief overview of the faceted search system, which comprises of two phases (see Fig. 2), first, facets and values generation, and second, faceted search. In the first phase, facets, which is a set of keywords resembling categories or aspects that helps in navigation/ refinement of search process is defined, followed by value extraction. In order to cater for different set of facets to suit domain’s need, the *facet definition* process first requires the definition of facets which deem suitable for the needs of a particular domain. A facet is specified as keyword, together with possible variants, such as <framework, frameworks>. These facets are stored as facet lexicon. Given the facet lexicon, the facet value extraction module extracts keywords which are related to each facet. In our context, facet values are instances related to a facet. E.g. for the facet “algorithm”, related values are like “Naive Bayes”, “Gaussian Naive Bayes”, “k-Means” etc. In this paper, two approaches are used to extract possible values given a facet.

i) Neighbor words detection – given a facet, the extraction algorithm performs checking on the neighbor words prior to the facet’s position. A halting criterion is imposed based on a stop word list like connective words. Extracted candidates are then weighted based on additional factors like capitalization, conceptual words in order to be selected as value. This approach is straight forward extraction based on facet that has been stated explicitly.

ii) Semi-supervised learning for non-neighbor words detection – for detection of values where its facet is not explicitly specified, a semi-supervised learning approach

is adopted to learn the possible values that have been associated with a facet (label) during step i). The learnt classifier is then used to label new values with the most likely facet for the values.

In the second phase, faceted search process receives the search query from user. The query goes through the information retrieval process to find the matched document in query-document matching. For faceted search, facets-values will be generated based the search results obtained from query-document matching module. The facet and values list produced will be displayed to users. In the facet selection module, user can choose a facet or a facet's value. The selection is then used as the input query again to refine the search process

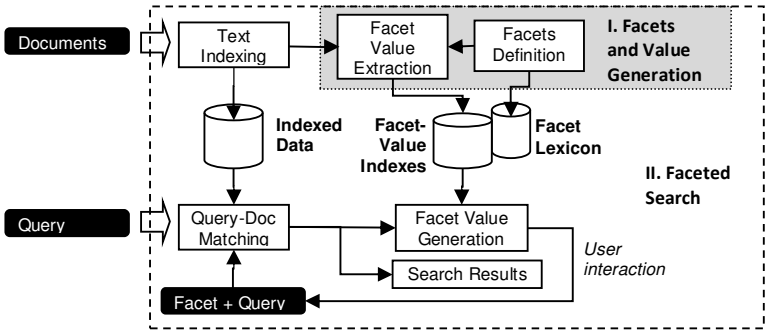

Fig. 2: The framework of faceted search system with flexible facets generation.

3 Demo

The system will be demonstrated in the context of bibliographical search which is accessible at http://ir.cs.usm.my/exsearch2/.

Acknowledgments. This research was supported by USM Research University Grant (1001/ PKOMP/811335: Mining Unstructured Web Data for Tour Itineraries Construction), Universiti Sains Malaysia.

References

1. Kules, B., Capra, R., Banta, M., Sierra, T.: What do exploratory searchers look at in a faceted search interface?, In Proc. of the 9th ACM/IEEE-CS joint conf. on Digital libraries, pp. 313-322. ACM (2009)
2. Tunkelang, D.: Faceted Search. Morgan & Claypool Pubs. (2009)
3. Koren, J., Zhang, Y., Liu, X.: Personalized interactive faceted search. In Proc. of the 17th international conference on World Wide Web, pp. 477-486, ACM (2008)
4. Faceted DBLP, http://dblp.l3s.de (Last accessed: 2016)
5. Du, J., Jin, P., Zheng, L., Wan, S., Yue, L.: DBLP-Filter: Effectively Search on the DBLP Bibliography, In Proc. of WWW 2014 (WWW '14 Companion), ACM (2014)
6 Komamizu, T., Amagasa, T., & Kitagawa, H.: Facet-value Extraction Scheme from Textual Contents in XML Data. IJWIS, 11(3), 270-290 (2015)
7 Arenas, A., Grau, B. C., Kharlamov, E., Marciuska, S., Zheleznyakov, D., Jimenez-Ruiz, E.: 2014. SemFacet: semantic faceted search over yago. In Proceedings of WWW2014 (WWW '14 Companion), pp. 123-126, ACM (2014)
8 Sacco, G. M.: Exploratory Access to Wikipedia through Faceted Dynamic Taxonomies, In Proc. of the 9th Int. AAAI Conference on Web and Social Media, Workshop: Wikipedia, a Social Pedia: Research Challenges and Opportunities (2015)