DFT-Extractor: A System to Extract Domain-specific Faceted Taxonomies from Wikipedia

Bifan Wei [+] Jun Liu [+] Jian Ma [+] Qinghua Zheng [+] Wei Zhang [++] Boqin Feng [+]
[+] MOE KLINNS Lab, Department of Computer Science, Xi'an Jiaotong University, Shaanxi, China
[++] Amazon.com Inc, Seattle, WA 98109, USA
weibifan@gmail.com liukeen@mail.xjtu.edu.cn majianxjtu@gmail.com qhzheng@mail.xjtu.edu.cn
wzhan@amazon.com bqfeng@mail.xjtu.edu.cn

ABSTRACT
Extracting faceted taxonomies from the Web has received increasing attention in recent years from the web mining community. We demonstrate in this study a novel system called DFT-Extractor, which automatically constructs domain-specific faceted taxonomies from Wikipedia in three steps: 1) It crawls domain terms from Wikipedia by using a modified topical crawler. 2) Then it exploits a classification model to extract hyponym relations with the use of motif-based features. 3) Finally, it constructs a faceted taxonomy by applying a community detection algorithm and a group of heuristic rules. DFT-Extractor also provides a graphical user interface to visualize the learned hyponym relations and the tree structure of taxonomies.

Categories and Subject Descriptors
[Information systems]: World Wide Web---Web mining; [Computing methodologies]: Artificial intelligence---Natural language processing---Information extraction; [Networks]: Network properties---Network structure;

Keywords
Faceted taxonomy, Network motif, Wikipedia.

1. INTRODUCTION
A faceted taxonomy is a set of taxonomies called facets. Each facet is a set of terms structured by hyponym relations [1]. A domain-specific faceted taxonomy characterizes the taxonomies of a specific domain and is often utilized to manage taxonomic knowledge within this domain. Figure 1 shows two facets of the faceted taxonomy that is data structure specific. Domain-specific faceted taxonomies play an important role in faceted search [2], domain-specific search, domain knowledge construction and other knowledge-based intelligent systems.

The manual construction of faceted taxonomies by domain experts is time consuming, labor intensive, and poorly scalable. Furthermore, this approach is problematic in that the constructed taxonomies tend to be biased according to experts' view. How to effectively and automatically construct unbiased domain-specific faceted taxonomies is a challenging issue.

The preliminary study of automatic faceted taxonomy construction has been conducted by a few researchers [3, 4]. However, the extracted faceted taxonomies only contain generic entities, such as *time*, *price*, and *size*, but lacking domain terms. Recent research has focused on taxonomy construction from Web pages, especially from massive user-generated content, e.g., Wikipedia. Probase [5] learns a universal, probabilistic taxonomy containing 2.7 million concepts harnessed automatically from a corpus of 1.68 billion Web pages. Recent efforts on learning taxonomies from Wikipedia include MENTA [6], BabelNet [7] and WikiNet [8]. MENTA extracts large amounts of facts about entities, whereas BabelNet and WikiNet focus on explicit semantic relations harvested from the Wikipedia Category System (WCS). All of these works mainly rely on text features to extract taxonomies.

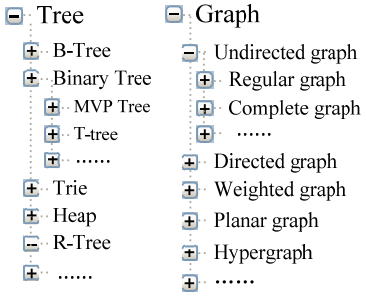

Figure 1. Two facets from the *Data-structure* faceted taxonomy

We developed a novel system called Domain-specific Faceted Taxonomy Extractor (DFT-Extractor[1]) that is different from the above approaches. With this system, we aim to learn faceted taxonomies from the topological structure of hyperlinks among Wikipedia article pages. The following are the key features of DFT-Extractor:

1) DFT-Extractor discovers hyponym relations based on the local connectivity patterns of a Wikipedia article graph, in which nodes and edges respectively represent article pages and hyperlinks. We found that hyponym hyperlinks are more likely to appear in some types of motifs (see Section 2.1 for the definition of *motif*) than in other types [9]. Thus, the different motifs of a hyperlink are utilized as discriminative features to identify the semantic relation of this hyperlink. These motif-based features can be directly used by a classifier. Good performance can be achieved without further combining text features. Structured information such as Wikipedia navigation boxes and category system is exploited by DFT-Extractor to label training sets heuristically.

[1] DFT-Extractor can be accessed from https://github.com/weibifan/DFT-Extractor/.

WWW 2013 Companion, May 13-17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2) DFT-Extractor utilizes a community detection algorithm to identify facets from Wikipedia article pages. It then uses a set of heuristic rules to construct taxonomies from the learned hyponym relations.

3) DFT-Extractor provides a graphical user interface to visualize the learned hyponym relations and the tree structure of taxonomies.

The rest of this demonstration is organized as follows. Section 2 shows the architecture of DFT-Extractor, explains important technical issues, and demonstrates the interactive visualization user interface. Section 3 reports a preliminary analysis on three data sets. Section 4 summarizes the conclusions of this demonstration.

2. DFT-EXTRACTOR SYSTEM

2.1 Preliminary

We first describe two important concepts used in this study.

Network motif: A network motif of a directed graph G is a sub-graph of G that occurs far more often in G than in randomized networks with the same degree distribution [10]. Table 1 shows that three-node network motifs consist of 13 nonisomorphic connection patterns. We refer to the jth motif in this table as motif j $(1 \leq j \leq 13)$ in the following sections.

Table 1. 13 types of three-node motifs

ID	Shape	ID	Shape	ID	Shape	ID	Shape
1		2		3		4	
5		6		7		8	
9		10		11		12	
13							

Domain-specific Faceted Taxonomy: A domain-specific faceted taxonomy (DFT) is a set of domain-specific taxonomies. It can be formally defined as $DFT = \{DT_i\}_n$. DT_i is a taxonomy and can be further defined as a triple $(V_i, R_i, root_i)$, where V_i is a set of domain taxonomic terms, $R_i \subset V_i \times V_i$ is a set of hyponym relations, and $root_i \in V_i$ is the only term that has no parents. The terms in different DT_i are preferably orthogonal and satisfy $\forall i, j \in [1..n], i \neq j, V_i \cap V_j = \varnothing$.

With the taxonomy *Tree* shown in Figure 1 as an example, $V_{Tree} = \{$*Tree*, *B-tree*, *Binary tree*, *Trie*, *Heap*, *R-tree*, ... $\}$, $R_{Tree} = \{$(*B-tree*, *Tree*), (*Binary tree*, *Tree*), (*Trie*, *Tree*), (*Heap*, *Tree*), (*R-tree*, *Tree*), (*MVP tree*, *Binary tree*), ... $\}$, and $root_{Tree} = Tree$.

2.2 Architecture of DFT-Extractor

Figure 2 shows that DFT-Extractor constructs faceted taxonomies employing a three-step framework. The function and characteristics of each module are described as follows:

Module I crawls the Wikipedia article pages and category pages of a given domain. With the domain *Data structure* as an example, this module crawls the article pages by traveling through article-article hyperlinks from a start position[2]; the module also crawls the category pages by traveling through category-category hyperlinks from another start position[3]. A set of URL regular expressions is used to filter out irrelevant article pages from the crawled pages, such as *External links*, *Languages*, and *View history*.

Module II exploits the local connectivity patterns of the Wikipedia article graph to discover hyponym relations among domain terms. The design of this module [9] enables it to automatically construct motif-based features from a Wikipedia article graph. However, the classification model learned from the motif-based features of one domain cannot be directly applied to another domain. Thus, this module automatically labels the training sets of different domains with the use of WCS and other structured information in Wikipedia [9].

Module III utilizes the learned hyponym relations of a domain to construct a coherent faceted taxonomy, in which each taxonomy is a strict tree structure.

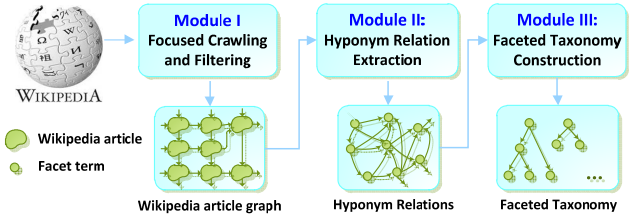

Figure 2. Architecture of DFT-Extractor

Module I is just a topical crawler specializing in Wikipedia, so the detailed implementation of the crawling algorithm will not be described further. In the following section, we will only focus on the description of modules II and III.

2.3 Hyponym Relation Extraction

Module II extracts hyponym relations from the Wikipedia article graph with the use of motif-based features. The extraction of hyponym relation can be further divided into three steps [9].

First, DFT-Extractor discovers all instances of three-node motifs from a domain-specific Wikipedia article graph. Then it enumerates all hyperlinks in every motif instance automatically.

Second, DFT-Extractor analyzes every hyperlink. This step is necessary because a hyperlink may simultaneously appear in multiple instances of a network motif. For every hyperlink, DFT-Extractor calculates its occurrences in different network motifs. Notation $O(i, j)$ is used to indicate how many times hyperlink i appears in the instances of motif j shown in Table 1; the notation implies the effect of motif j on hyperlink i.

Third, DFT-Extractor builds a feature vector for every hyperlink within the Wikipedia article graph. This vector is based on the occurrences of different network motifs. The hyperlink occurrences of network motifs have varying significance, which can be weighted by the Z-Scores of corresponding network motifs; thus, the weighted vector component for hyperlink i and motif j can be formalized as follows:

$$F(i, j) = O(i, j) * \text{Z-Score}(j) \quad (1)$$

[2] http://en.wikipedia.org/wiki/Data_structure
[3] http://en.wikipedia.org/wiki/Category:Data_structures

Z-Score() in Formula (1) is defined as $(N(j) - \overline{N_r(j)})/\sigma_r(j)$ [10], where $N(j)$ is the number of occurrences of motif j $(1 \leq j \leq 13)$ in network G. $\overline{N_r(j)}$ is the average number of occurrences of motif j in an ensemble of randomized networks with the same degree of distribution as network G. $\sigma_r(j)$ is the standard deviation of $N_r(j)$.

Moreover, the feature vector of a hyperlink i is represented as follows:

$$FV(i) = (F(i,1), F(i,2), \cdots, F(i,13)) \qquad (2)$$

With these motif-based features, the module converts the hyponym relation extraction problem into a binary classification problem.

The classification models learned from a training set can perform fairly well in the test set of the same domain. However, these models cannot be directly applied to another domain for hyponym relation discovery. For this reason, the category pages in WCS and navigation boxes within the Wikipedia article pages are leveraged to label the training sets of different domains with minimal human involvement. This module automatically extracts the structured information from Wikipedia pages to heuristically label training sets. This makes DFT-Extractor robust to domain changes.

The articles and learned hyponym relations among them form a directed graph $HG(V, E)$, where V is a set of article pages and E is a set of hyponym relations.

Figure 3 shows a screenshot of the module II user interface that illustrates the procedure and the results of hyponym relation extraction.

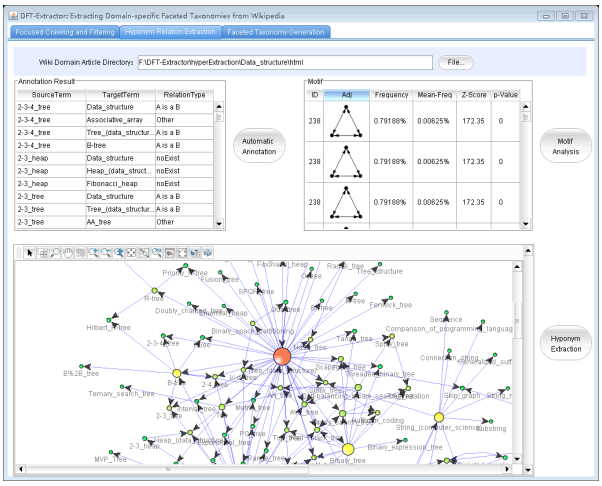

Figure 3. Screenshot of module II

2.4 Faceted Taxonomy Generation

Module III automatically induces a faceted taxonomy $DFT = \{DT_i\}_n$ over $HG(V, E)$.

First, the module discovers facets from the Wikipedia article graph by applying the Louvain community detection algorithm [11], where the resolution parameter is set to the default value 1. Every community, which is a subgraph of the Wikipedia article graph, corresponds to a facet. The article pages within a community correspond to the candidate terms of this facet.

Then the module identifies the root of the taxonomy from every community by using degree centrality. A high node degree indicates that this node has more connections with other nodes and is more likely to be a root node. Thus, all nodes in a community are ranked by degree centrality, and the top-2 nodes are selected as reference nodes. DFT-Extractor exploits the following two rules to infer the root of a taxonomy.

Rule 1: If the two reference nodes have common parent nodes in this community, then the highest common ancestor is selected as the root of this facet.

Rule 2: If the two reference nodes have no common parent node, then the highest ancestor of the reference node with a higher degree is selected as the root of this facet.

Third, this module constructs a taxonomy for every facet from the root and candidate terms. With the root of a taxonomy taken as the start point, the module travels through the learned hyponym relations in a breadth-first order. Thus, the traveled terms and hyponym relations among them constitute the taxonomy of a facet.

Figure 4 shows a screenshot of the module III user interface. The upper part visualizes the results of community detection. The size of each node in a community is proportional to its degree centrality. The lower part of the screenshot displays the extracted faceted taxonomy in tree view.

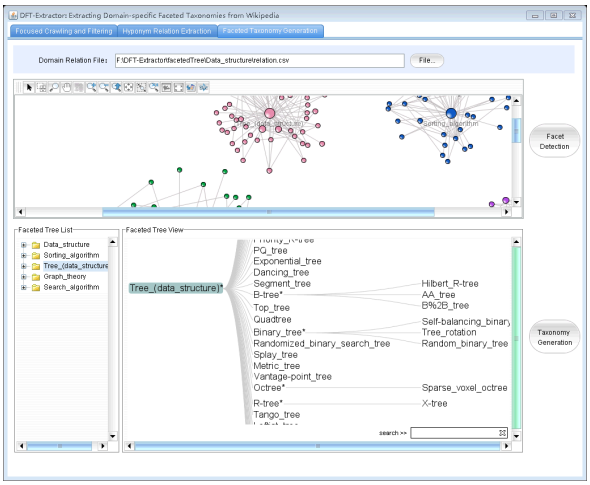

Figure 4. Screenshot of module III

2.5 Implementation Details

DFT-Extractor is designed to be easy to use and is written in Java for cross-platform support. The user interface of DFT-Extractor contains three tab panels, each corresponding to a module shown in Figure 1.

To improve robustness and scalability, the Java Universal Network/Graph framework (JUNG) [12] and the Prefuse visualization toolkit [13] are incorporated into DFT-Extractor for graph/network analysis and visualization.

3. PRELIMINARY RESULTS

From our preliminary experiments based on three data sets (*Data Structure*, *Data Mining*, and *Computer Network*), we found that *SVM* and *Random Forest* (*RF*) perform fairly well with the use of motif-based features [9]. DFT-Extractor adopts these two popular classification algorithms to extract hyponym relation. In the three data sets, 30% of all the hyperlinks are used as the training set labeled with structured information. The other 70% are used as the test set.

The hyponym relation extraction based on lexico-syntactic pattern matching is very popular and effective in processing Web pages.

Therefore, the baseline system utilizes 10 lexico-syntactic patterns with the highest coverage to extract hyponym relations. Table 2 shows the performance of the baseline and DFT-Extractor with the use of the two classification algorithms. The best F1 scores are in bold.

Table 2. Experimental results based on the three data sets

Data sets	Baseline			DFT-Extractor +SVM			DFT-Extractor +RF		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Data Structure	0.487	0.358	0.413	0.852	0.524	0.649	0.811	0.652	**0.723**
Data mining	0.402	0.334	0.365	0.880	0.529	0.661	0.803	0.608	0.692
Computer Network	0.584	0.326	**0.418**	**0.977**	0.513	**0.673**	0.814	0.534	0.645

Table 2 also shows that 1) the motif-based features are effective for hyponym relation extraction. 2) The F1 scores of motif-based classification algorithms are also higher than those of the lexico-syntactic based baseline.

Table 3 shows the number of facets and the corresponding roots of taxonomies extracted by DFT-Extractor.

Table 3. Facets of the three data sets

Data sets	# Facets	Roots of Taxonomies
Data Structure	4	Graph, Tree, Array, Computer algorithm
Data mining	5	Data, Statistical classification, Cluster analysis, Regression analysis, Association rule learning
Computer Network	5	Internet, LAN, Networking hardware, OSI model, Protocol

Figure 5 visualizes the taxonomy *Computer algorithm* of the *Data Structure* domain.

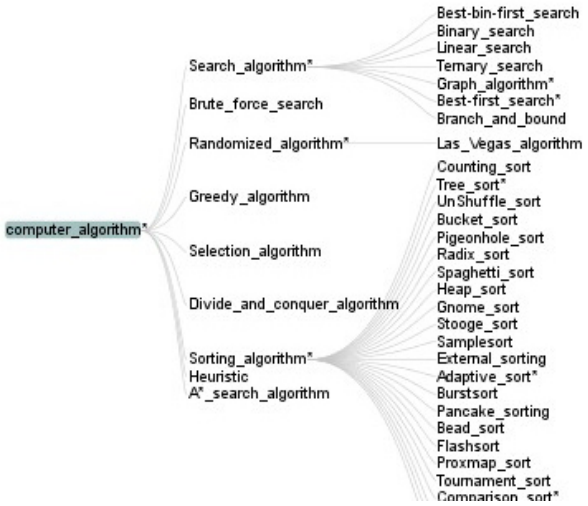

Figure 5. Taxonomy *Computer algorithm* of domain *data structure*

4. CONCLUSION AND FUTURE WORK

We have developed an easy-to-use system called DFT-Extractor. This system is capable of automatically extracting domain-specific faceted taxonomies from Wikipedia with little domain expert supervision. Most notably, this is the first system to our knowledge that discovers hyponym relations with the use of motif-based features. Our experimental results provide compelling evidence that the three-step framework that DFT-Extractor adopts is effective in addressing the faceted taxonomies extraction problem that is domain-specific. However, some limitations of DFT-Extractor are worth noting. For example, only three-node motifs are analyzed for hyponym relation extraction without adding text features. The future extension of DFT-Extractor should therefore include improving the performance of hyponym relation extraction. This improvement can be achieved by combination of text features and development of new algorithms to process complex connectivity patterns.

5. ACKNOWLEDGMENT

The research was supported in part by the National Science Foundation of China under Grant Nos. 61173112, 61221063 and 61202184; National High Technology Research and Development Program of China under Grant No. 2012AA011003; National Key Technology R&D Program under Grant No. 2011BAK08B02; Cheung Kong Scholar's Program.

6. REFERENCES

[1] Y. Tzitzikas, N. Spyratos, P. Constantopoulos, and A. Analyti. Extended faceted taxonomies for web catalogs. In *Proc. of WISE-02*, pages 192-204, 2002.

[2] B. Wei, J. Liu, Q. Zheng, W. Zhang, X. Fu, and B. Feng. A survey of faceted search. *Journal of Web engineering,* vol. 12, pages 041-064, 2013.

[3] W. Dakka and P. G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. In *Proc. of ICDM-08*, pages 466-475, 2008.

[4] E. Stoica, M. A. Hearst, and M. Richardson. Automating creation of hierarchical faceted metadata structures. In *Proc. of HLT-NAACL-07*, pages 244-251, 2007.

[5] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: a probabilistic taxonomy for text understanding. In *Proc. of SIGMOD-12*, pages 481-492, 2012.

[6] G. d. Melo and G. Weikum. MENTA: inducing multilingual taxonomies from Wikipedia. In *Proc. of CIKM-10*, pages 1099-1108, 2010.

[7] R. Navigli and S. P. Ponzetto. BabelNet: building a very large multilingual semantic network. In *Proc. of ACL-10*, pages 216-225, 2010.

[8] V. Nastase, M. Strube, B. Boerschinger, C. Zirn, and A. Elghafari. WikiNet: a very large scale multi-lingual concept network. In *Proc. of LREC-10*, pages 1015-1022, 2010.

[9] B. Wei, J. Liu, J. Ma, Q. Zheng, W. Zhang, and B. Feng. MOTIF-RE: motif-based hypernym/hyponym relation extraction from wikipedia links. In *Proc. of ICONIP-12*, pages 610-619, 2012.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science,* vol. 298, pages 824-827, 2002.

[11] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, pages P10008, 2008.

[12] J. Madadhain, D. Fisher, P. Smyth, S. White, and Y. B. Boey. Analysis and visualization of network data using JUNG. *Journal of Statistical Software,* vol. 10, pages 1-35, 2005.

[13] J. Heer, S. K. Card, and J. A. Landay. Prefuse: a toolkit for interactive information visualization. In *Proc. of SIGCHI on Human Factors in Computing Systems*, pages 421-430, 2005.