Extraction of feature words with the same generality level as query using restricted bootstrapping

Zeng, Jun
Graduate School of Information Science, Kyushu University

Sakai, Toshihiko
Graduate School of Information Science, Kyushu University

Flanagan, Brendan
Graduate School of Information Science, Kyushu University

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings of the 2012 IEEE 14th International Conference on Commerce and Enterprise Computing, CEC 2012, pp.171-176, 2012-12-01
バージョン：
権利関係：

Extraction of Feature Words with the Same Generality Level as Query using Restricted Bootstrapping

Jun Zeng
Graduate School of Information Science and Electrical Engineering,
Kyushu University,
Fukuoka, Japan
zeng.j.000@s.kyushu-u.ac.jp

Toshihiko Sakai
Graduate School of Information Science and Electrical Engineering,
Kyushu University,
Fukuoka, Japan
2IE11060Y@s.kyushu-u.ac.jp

Brendan Flanagan
Graduate School of Information Science and Electrical Engineering,
Kyushu University,
Fukuoka, Japan
bflanagan.kyudai@gmail.com

Sachio Hirokawa
Research Institute for Information Technology,
Kyushu University,
Fukuoka, Japan
hirokawa@cc.kyushu-u.ac.jp

***Abstract*—It is not so simple to get an appropriated level of search result among a large number of targets which might be too general or too specific. Hints for the query are valuable for a user to expand or shrink his next search step, if the hints would be shown with their levels compared with the user's original query. This paper proposes a method to extract feature words with the same level as user's query using restricted bootstrap. Examples are shown to demonstrate the effectiveness of the method on tourism blogs. The paper proposes an evaluation measure for the similarity of levels for words based on WordNet.**

Keywords-restricted bootstrapping; generality level; feature word; WordNet

I. INTRODUCTION

The Web, as the largest database, often contains information that may be interesting for researchers and the general public. The quantity of information available today is more than at any other point in history, but with this wealth of information comes even greater challenges. Due to the popularization of web search engines, finding information has become an easy exercise by just typing a few keywords into a search engine text-box. However, search engines always return large amounts of search result, which users need to spend considerable time reviewing to find the information he/she wants. Because of this, most search engines employ methods to rank the results to provide the "best" results first. In other words, the usefulness of a search engine depends on the relevance of the result set it gives back.

Nevertheless, ranking search result has some intrinsic shortages. On the one hand, most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the practice of allowing advertisers to pay money to have their listings ranked higher in search results. The top-ranking search results sometimes have little relevance to user's query. On the other hand, although the ranking algorithms have become more and more mature, they still have some insuperable problems. The search results are displayed as linearly, and the hierarchy of the search results is ignored. It is difficult to get an appropriated level of search result among a large number of targets which might be too general or too specific.

For solving the problems mentioned above, we change another viewpoint to return search results. We consider that the web pages have a generality level. The levels of this generality are considered to be hierarchical and scored according to the topics of the web pages. For example, a page about animal has a higher level than a page about mammal, while a page about dog has a lower level than the mammal page. We determine the generality level of a page by analyzing the generality level of words in the page. The more high-level words the page contains, the higher the generality level the page will have.

We are considering sorting web pages according to their generality level. When a user conducts a search, only the search results which have the same generality level as the user's query will be displayed, rather than simply ranking the search results. By comparing the generality level between query and results, users can easily adjust their keywords to obtain better results. In order to obtain the same generality level pages, we must get words with the same generality level first. In this paper, we propose a method to extract the features words with the same generality level as user's query from web pages using restricted bootstrapping. Bootstrapping [1] is a technique used to iteratively improve a classifier's performance and extract

information from documents. WordNet [2] is used to define the generality level of words. A prototype system is also developed. We collected 10,000 tourism blog pages as experiment data and extracted the feature words with same generality level from the tourism blogs.

This paper is structured as follows: In Section II we briefly introduce the related works. In Section III we define words with the same generality level and introduce the bootstrapping method to extract words with the same generality level. In section IV, we introduce the evaluation experiment. Finally, Section V describes our conclusions and proposed future work.

II. RELATED WORKS

A. *Extraction of Words or Phrases from Web Pages*

There is a lot of related work on the extraction of Words or Phrases from web pages.

P. Turney [3] proposed a method for automatically extracting keyphrases which are the keywords appearing on the first page of each academic journal article from text as a supervised learning task. The keyphrases They are treating a document as a set of phrases, which the learning algorithm must learn to classify as positive or negative examples of keyphrases.

Huang Y.-F [4] proposed a general framework that could automatically extract key-phrases from a collection of web pages concerning a specific topic with the help of The Free Dictionary and then constructed a personal knowledge base. Both the base and visual feature in a web page are used to calculate the weight of each candidate phrase. The system extracts top p% key-phrases for each web page based on these two features and then generates a term set using union operators. Next, the system builds the relationships between terms in the term set by referencing The Free Dictionary, and then generates a list of terms sorted by weights. With the top q terms specified by users, a semantic graph can be constructed to present the part of a personal knowledge base, which shows the relationships between terms from the same domain.

B. *Bootstrapping for Information Extracion*

Bootstrapping is a technique used to iteratively improve a classifier's performance and extract information from documents. Bootstrapping can collect required information step by step with just few words at first.

Fang Tian et al. [5] proposed an effective method to automatically extract hyponym from the Web for Chinese. The method extracts hyponyms for a given hypernym through weak supervision in two stages: the first stage is submitting a hypernym and a seed hyponym as a query to Web search engine, and automatically extracting hyponyms matching with a Chinese doubly anchored hyponymy pattern from the Web by bootstrapping.

M. Komachi et al. [6] demonstrated the semantic drift of Espresso-style bootstrapping has the same root as the topic drift of Kleinberg's HITS, using a simplified graph-based reformulation of bootstrapping. They confirmed that two graph-based algorithms, the von Neumann kernels and the regularized Laplacian, can reduce the effect of semantic drift in the task of word sense disambiguation (WSD) on Senseval-3 English Lexical Sample Task. They also proposed a algorithm achieve superior performance to Espresso and previous graph-based WSD methods, even though the proposed algorithm has less parameters and are easy to calibrate.

Although the bootstrapping method is known as an application of the Page-rank technique for documents and words. However, the technique calculates the score of the words by mutually propagating the score of the words and the documents. However, sometimes the result is far away from the initial query word. The problem is known as topic drift. S. Hirokawa [7] proposed to restrict the words to be to the top t words in the process of bootstrapping. The method is simpler than the technique known so far. The method is applied for the real bankruptcy information documents to extract the bankruptcy causes strongly related to the query. It is confirmed that the method prevents the topic drift. In this paper, restricted bootstrapping is used to extract the feature words with the same generality level as query from web pages

III. DEFINITION OF FEATURE WORDS WITH THE SAME GENERALITY LEVEL

A. *Generality Level in WordNet*

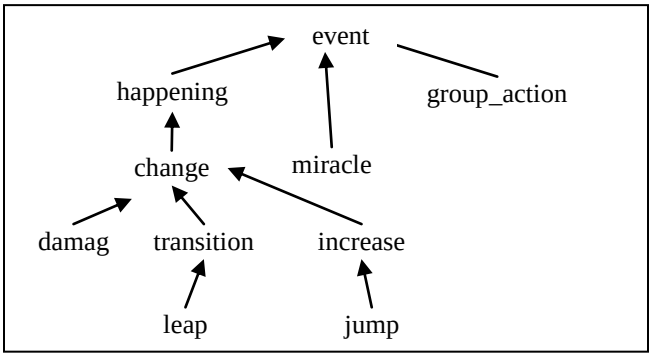

Figure 1. A small subgraph of the WordNet graph

In this paper, WordNet is used to define the words with the same generality level. WordNet is a semantic lexicon for the English language that is used extensively by computational linguists and cognitive scientists. In 2009, Japanese WordNet released. WordNet groups words into sets of synonyms called synsets and describes semantic relationships between them. One such relationship is the is-a relationship, which connects a hyponym (more specific synset) to a hypernym (more general synset). For example, a plant organ is a hypernym to plant root and plant root is a hypernym to carrot. Both nouns and verbs are organized into hierarchies defined by hypernym. The hyponym and hypemym form the generality levels in WordNet graph. Figure 1 shows the generality level in a small subgraph of the WordNet graph.

The graph is directed and acyclic, though not necessarily a tree since a synset can have several hypernyms. In the graph, each vertex v is an integer that represents a synset, and each directed edge $v \rightarrow w$ represents that w is a hypernym of v.

B. *Definition of Feature Words with the Same Generality Level*

This section will introduce the definition of words with the same generality level according the WordNet graph. As mentioned above, words are organized into hierarchies in WordNet. The feature words with the same generality level can be defined as following:

Define 1: Graph $G=(V,E)$ is directed and acyclic, $u \in V$, $v \in V$, $w \in V$, here w is a successor of both u and v. Path $P(u,w)=\{u_i \in V | < u,u_1 >, < u_1,u_2 > ... < u_m,w >\}$ leads from u to w and path $P(v,w)=\{v_i \in V | < v,v_1 >, < v_1,v_2 > ... < v_n,w >\}$ leads from v to w. If $m=n$ then we define u and v have the same level.

Define 2: In WordNet Graph, if word A and B have the same level then we define A and B are words with the same generality level.

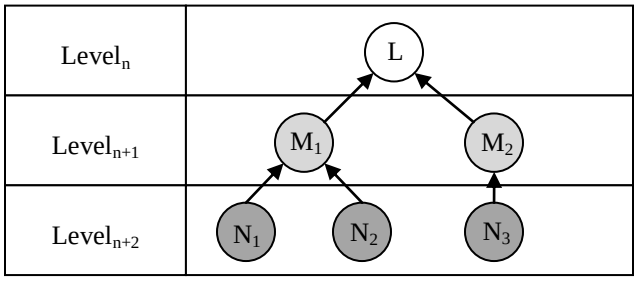

Figure 2. Txaonomic Levels of Words in WordNet Graph

Figure 2 shows an example of the generality level of words in WordNet Graph. There are three levels in the figure. M_1 and M_2 have the same generality level, and N_1, N_2 and N_3 have the same generality level. Our purpose is to extract such words with the same generality level from web pages.

IV. Extraction of Feature Words with the Same Generality Level using Restricted Bootstrapping

This section will introduce the restricted bootstrapping algorithm to extract the feature words with the same generality level. It is based on the index of documents and the index of sentences. In order to build the index of documents and the index of sentences "ChaSen[1]" and "GETA[2]" are used. "ChaSen" is a morphological parser for the Japanese language. "GETA" is a generic engine for transposable association. Figure 3 shows the procedure of restricted bootstrapping. There are 5 steps as following:

Step 1: Input a query (a keyword), and return the set of documents which contain the keyword;

Step 2: Get the set of sentences that are contained by the set of documents;

Step 3: Return the words that associate with the keyword. If the feature words satisfy a condition which can be determine in advance then go to step 5, otherwise go to step 4;

Step 4: Return the set of sentences which contain the one or more feature words, and then go to step 3;

Step 5: Return the feature words.

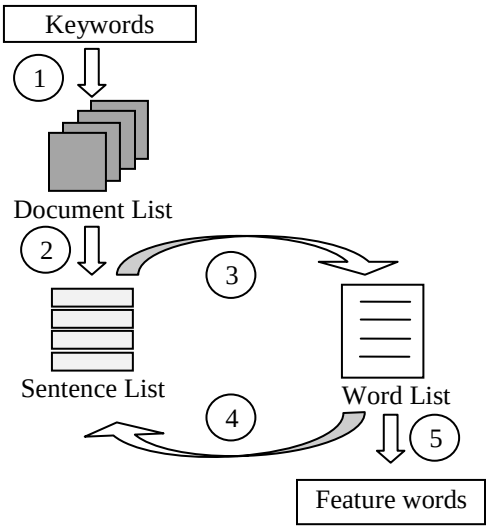

Figure 3. Procedure of Restricted Bootstrapping

```
main(w, t){
    input w: keyword, t: threshold
    output FW: ordered list of feature words
    D = Search-documents(w)
    # D is the set of documents that contain w
    S = Sentence(D)
    # the set of sentences in the set of document D
    T = Bootstrap(S, t)
    # the set of sentences obtained by the algorithm;
    FW = Top-keywords(T, t) # top t words in T
    return FW
}
Bootstrap(S, t){
    input S: set of sentences, t: threshold
    output S:set of sentences
    W = null
    # initiate the set of top words W
    for(i=0; i<max iteration; i++){
        W = Top-keywords(S, t) # top t words in S
        S = Search-sentences(W)
        # the set of sentences containing one or more words
        # in W
    }
    return S
}
```

Figure 4. Algorithm of Restricted Bootstrapping

Figure 4 shows the algorithm of restricted bootstrapping in detail. The algorithm constructed from the following four functions:

Search-documents(*w*) returns the set of documents which contain the keyword *w*.

Sentence(*D*) returns the set of sentences in the set of documents *D*.

Search-sentence(*W*) returns the set of sentences which contain one or more words in the set of words *W*.

Top-keywords(*S, t*) returns top t words, which are sorted automatically by "GETA", in the set of sentence *S*.

V. CONCLUSION AND FUTURE WORKS

REFERENCES

[1] Z. Mooney and4 D. Duval, Bootstrapping: A nonparametric approach to statistical inference, 1993.

[2] George A. Miller and C. Fellbaum, "WordNet then and now", Language Resources and Evaluation, Vol. 41(2), 2007, pp. 209-214.

[3] P. Turney, "Learning Algorithms for Keyphrase Extraction", Information Retrieval Journal, 2(4) , 2000, pp.303-336.

[4] Huang Y.-F, and Ciou C.-S., "Constructing personal knowledge base: Automatic key-phrase extraction from multiple-domain web pages", 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2011, Shenzhen, 24-27 May 2011. pp. 65-76.

[5] Tian, F.ab, Yuan, C.c, and Ren, F, "Hyponym extraction from the web by bootstrapping", IEEJ Transactions on Electrical and Electronic Engineering, Vol. 7(1), 2012, pp.62-68

[6] M. Komachi, T. Kudo, M. Shimbo, Y. Matsumoto, Semantic Drift in Espresso-style Bootstrapping: Graph-theoretic Analysis and Evaluation in Word Sense Disambiguation, Journal of JSAI, Vol.25, No.2, pp. 233-242 (in Japanese)

[7] S. Hirokawa, "Feature Extraction Using Restricted Bootstrapping", The 11th International Symposium on Innovative E-Services and Information Systems (IEIS 2012), Shanghai, China, May 30-31, 2012, pp. 283-288.