Autopedia: Automatic Domain-Independent Wikipedia Article Generation

Conglei Yao
HP Labs China
HP Labs
Beijing, China
conglei.yao@hp.com

Xu Jia
Dept. of Comput. Sci.
Renmin Univ. of China
Beijing, China
jiaxu@ruc.edu.cn

Sicong Shou
Dept. of Comput. Sci.
Peking Univ.
Beijing, China
ssc@net.pku.edu.cn

Shicong Feng
HP Labs China
HP Labs
Beijing, China
shicong.feng@hp.com

Feng Zhou
School of Comput. Sci.
Beijing Univ. of Technology
Beijing, China
feng.zhou2@hp.com

HongYan Liu
School of Economics and Management
Tsinghua Univ.
Beijing, China
liuhy@sem.tsinghua.edu.cn

ABSTRACT

This paper proposes a general framework, named *Autopedia*, to generate high-quality wikipedia articles for given concepts in any domains, by automatically selecting the best wikipedia template consisting the sub-topics to organize the article for the input concept. Experimental results on 4,526 concepts validate the effectiveness of Autopedia, and the wikipedia template selection approach which takes into account both the template quality and the semantic relatedness between the input concept and its sibling concepts, performs the best.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge acquisition*

General Terms

Algorithms, Experimentation

Keywords

Wikipedia, article generation, template selection, domain independent

1. INTRODUCTION

As the largest online encyclopedia, Wikipedia provides a large number of human-edited articles for popular concepts in most domains. In each article, some sub-topics are selected and described to provide a summary of the corresponding concept. Although Wikipedia has provided a large number of high-quality articles for popular concepts, it only covers a small part of popular concepts in major domains due to the huge cost of manually creating and editing articles.

Our objective is to automatically generate high-quality Wikipedia articles for any given concepts in any domains.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

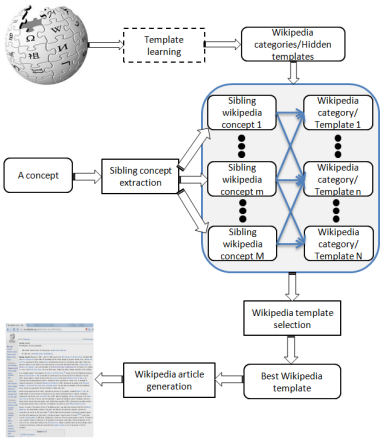

Figure 1: Framework of automatic domain-independent wikipedia article generation

To achieve this goal, we must discover the best template for each given concept to characterize its major sub-topics which should be covered by its article, as well as the related keywords for each sub-topic. Then, we can retrieve the related web pages and extract the corresponding paragraphs for each sub-topic, to generate the article.

We propose a general framework to generate high-quality Wikipedia articles for given concepts in any domains. Under this framework,we propose four ever-increasing methods to select the best template automatically, to support the domain-independent wikipedia article generation.

2. AUTOPEDIA FRAMEWORK

We propose the framework of Autopedia as illustrated in Figure 1. The offline component, *template learning*, learns the hidden template for each Wikipdia category by clustering the sub-topics in the articles of this category. After learning the templates, for a given concept, we can generate its Wikipedia article in three steps.

In the first step, *sibling concept extraction*, we can get a

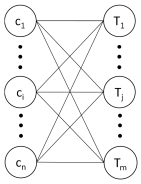

Figure 2: Bipartite graph between sibling concepts and wikipedia templates

set of Wikipedia concepts which are of the same classes with the given concept by leveraging the automatic set instance extraction tools (e.g., Google Sets).

After that, for each sibling concept, we can get its categories in Wikipedia, and establish the relation between it and the hidden templates of these categories. We then propose an approach (in *wikipedia template selection* component) to analyze such relationships to calculate the importances of templates, and select the best template with the highest importance value.

Finally, in *wikipedia article generation* component, for each sub-topic in the selected template, we can utilize its keywords to create queries, to retrieve related paragraphs from the Web, and generate the final Wikipedia article.

2.1 Wikipedia template selection

We can represent the relationships between the sibling concepts and wikipedia templates using a bipartite graph G as illustrated in Figure 2, where c_i is a sibling concept, T_j is a wikipedia template. An edge exists between c_i and T_j when c_i belongs to the wikipedia category from whose articles the template T_j is learned.

Intuitively, we can use the indegree of a template T_j to measure its importance:

$$IM(T_j) = \frac{deg^-(T_j)}{\max_{k=1}^{m} deg^-(T_k)} \quad (1)$$

where $deg^-(T_j)$ is the indegree of T_j.

If we take into account the semantic relatedness between each sibling concept with the input concept, we can calculate the template importance as:

$$IM_{st}(T_j) = \frac{deg'^-(T_j)}{\max_{k=1}^{m} deg'^-(T_k)} \quad (2)$$

$$deg'^-(T_j) = \sum_{i=1}^{n} f(i,j) * r(c_i, c_0) \quad (3)$$

where $r(c_i, c_0)$ is the semantic relatedness between c_i and the input concept c_0, and can be measured using Pointwise Mutual Information-Information Retrieval (PMI-IR) [1] method. The value of $f(i,j)$ is 1 if an edge exists between c_i and T_j, and is 0 if not.

We can also integrate the quality of the each wikipedia template into the template importance calculation:

$$IM_{qt}(T_j) = IM(T_j)Q(T_j) \quad (4)$$

where $Q(T_j)$ is the quality score of T_j, and the details of $Q(T_j)$ can be found in the technical report [1].

[1]http://fusion.hpl.hp.com/autopedia/techreport.pdf

Table 1: Top 5 best categories and the corresponding concepts

Template/category name	Uncovered concepts
BRIT Award winners	Guy Masterson, Jo Human
BAFTA winners (people)	Albert Fortell, Daniel Breton
Italian Ministers of Foreign Affairs	Emilio Sereni, Guido Fanti
PlayStation 2 games	Operation Air Assault
Banks of Norway	Modum SpareBank, Totens Sparebank

Table 2: Performance comparison of the four proposed template selection approaches

Approach	TemplatePrecision
TS	0.2749
TS_{st}	0.2945
TS_{qt}	0.6799
TS_{cb}	**0.7611**

We can also integrate both the semantic relatedness between concepts and the template quality into the template importance calculation:

$$IM_{cb}(T_j) = IM_{st}(T_j)Q(T_j) \quad (5)$$

Through this way, we have four different template selection approaches, TS, TS_{st}, TS_{qt} and TS_{cb}, which correspond to the four template importance calculation equations, IM, IM_{st}, IM_{qt} and IM_{cb}, respectively.

3. EXPERIMENTS

we randomly select 4,526 concepts from 17,908,995 concepts which have been labeled as concepts in Wikipedia but have no corresponding articles, and use Google sets to extract sibling concepts for them.

We firstly manually annotate the best template for each selected concept, from the templates of its sibling wikipedia concepts. From the annotation results, we find that all the 4,526 concepts corresponding to 831 different best categories. We also list the top 5 best template/category names as well as one or two concepts in Table 1. We can see that the selected concepts can cover enough domains to validate the effectiveness of our proposed approaches.

We then use the proposed approaches to select the best templates for all the selected concepts, and compare the results with the gold standard to compute the corresponding *template precision* values. The results are illustrated in Table 2. We can see that the simple indegree based method TS can only get poorer performance, and better performance can be achieved by incorporating the semantic relatedness between selected concepts and their sibling concepts (TS_{st}), or the template quality scores (TS_{qt}). Furthermore, TS_{qt} can achieve much better performance than TS_{st}, which indicates that the template selection is more sensitive to template quality, and our template quality score calculation method is effective. Finally, we can see that the best performance is achieved by TS_{cb}, which combines both the template quality and the semantic relatedness between uncovered concepts and their sibling concepts.

4. REFERENCES

[1] P. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the twelfth european conference on machine learning (ecml-2001)*, pages 491–502, 2001.