Learning Web Queries for Retrieval of Relevant Information about an Entity in a Wikipedia Category

Vikrant Yadav
Indian Institute of Technology, Roorkee
India.
vikrantiitr1@gmail.com

Sandeep Kumar
Indian Institute of Technology, Roorkee
India.
sandeepkumargarg@gmail.com

ABSTRACT

In this paper, we present a novel method to obtain a set of most appropriate queries for retrieval of relevant information about an entity from the Web. Using the body text of existing articles in a Wikipedia category, we generate a set of queries capable of fetching the most relevant content for any entity belonging to that category. We find the common topics discussed in the articles of a category using Latent Semantic Analysis (LSA) and use them to formulate the queries. Using Long Short-Term Memory (LSTM) neural network, we reduce the number of queries by removing the less sensible ones and then select the best ones out of them. The experimental results show that the proposed method outperforms the baselines. Existing approaches are performing better in generation of the relevant section title queries by extraction from the headings of the Wikipedia articles as compared to the generation of queries by extraction from the body text of the articles. Whereas, the experimental results show that the proposed approach can perform equally well and even better in extraction of the relevant queries from the body text of the Wikipedia articles.

Keywords

Information retrieval; Query generation; Wikipedia article generation; Long Short-Term Memory Neural Networks (LSTM)

1. INTRODUCTION

Wikipedia contains detailed information about entities generally written in a coherent manner. However, there are still a large number of entities on which Wikipedia articles are not written. Thus, there is a need to create a system for automatic generation of Wikipedia articles. In past, Sauper et al. [1] proposed a structure aware approach for generating Wikipedia article automatically about any entity given its Wikipedia category.

Acquiring a set of queries for retrieval of important information has been studied by a large number of researchers in the past. However, these studies extract query patterns that contain stop-words like "is an", "was an", "of", etc. termed as generic patterns by Tanaka et al. [2]. These are not useful because search engines like Google, Yahoo tend not to retrieve and rank Web pages based on stop-words. Also, specific query patterns, like "is an opera by", "is a comic", "is the second" for category Opera contain more stop-words than the keywords. In fact many of the keywords do not reflect the type of content present in the articles of a category.

In this paper, we present a novel method to obtain a set of most appropriate queries for retrieval of relevant and important information about an entity from the Web.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
DOI: http://dx.doi.org/10.1145/2872518.2891114.

2. PROPOSED METHOD

An entity is a topic for which we are creating a Wikipedia article. A template is a set of section headings that represents the common structure of articles in a Wikipedia category. For example, (Early Life, Career, Television, Personal Life) is a possible template for American Male Actors category. We generate the template by selecting four most frequent headings present in the articles belonging to a particular category. A query to the Web search engine is of the format ("*X template_section query_term*") where *template_section* is the section of the template for which we are collecting information, and *query_term* is one of the queries obtained by our approach for the template section. X is the entity name. For example, query "Brad Pitt + career + films" indicates that "Brad Pitt" is the entity, "career" is one of the sections in the template for *American Male Actors* category and "films" is the query term obtained by our method. Given entity category (e.g., *American Male Actors*, *Cities and Towns*), the goal of this study is to acquire a set of queries for the retrieval of relevant information from the Web. The method consists of three steps – Query Extraction, Query Reduction and Best Query Selection.

2.1 Query Extraction

Queries generated from the body text of each section are more section-specific as they represent the type of information generally present in that section. We use topic modeling to find the most common topics talked about in the existing articles belonging to the same category for each heading of the template. For a section, we run a topic modeling technique, called Latent Semantic Analysis (LSA), on the contents of the same section in different articles to find out the top topics.

In the results of LSA for different sections, almost always the top one or two topics have far higher singular values than rest of the topics. Thus, we discard the rest of the topics, consider top 10 weighted terms in the top 2 topics and combine them to form possible query terms.

2.2 Query Reduction

The queries which make sense are the ones whose terms occur very close to each other in the sentences of existing articles. A good technique to measure the closeness between words in the text is LSTM neural network [3]. We trained different LSTM networks for each section in the articles of a category. The input sequence is a list of word embeddings, i.e. vector-based semantic representation of each word in the input sentence. We used Word2Vec [4] to get the word embedding for each input word. The output dimension of the network at each time-step is chosen same as the dimension of input word embedding, i.e. 300.

After training, for each possible bigram query, we feed the first term of the query as input to the network and calculate the cosine similarity between the output vector and the word embedding of the second term. This cosine similarity is the score of that query. Queries whose terms occur close to each other in the existing

articles have high score. We select the top 10 queries with similarity above a threshold of 0.80.

2.3 Best Query Selection

For each section in the template of a Wikipedia category, we create training and testing datasets of articles. For each query of the section learned from the training dataset, we query the Web using any of the popular search engines for every article in the testing dataset and retrieve the top 10 webpages.

An excerpt is the text inside the paragraph ( ) tags in these webpages. A query's score is defined as:

$$\text{query score, } q = \sum_{j=1}^{j=m} s_j \quad (1)$$

Where "m" is the total number of articles in testing dataset and "s_j" is the maximum cosine similarity score between any of the excerpt (paragraph) "i" retrieved by the query and the original text of the section in the Wikipedia article "j", i.e.

$$s_j = max(cosine\_sim(i, section\_body_j)), \forall excerpts\ i \quad (2)$$

The top 5 best scoring queries are selected and stored for each section of the template.

3. EXPERIMENTS AND RESULTS

3.1 Data

We used the articles of categories *American Male Actors* and *Cities and Towns in India*. We divide the article set of a given category into two subsets. One is the training set, used for finding and filtering the probable queries, and other is the testing set, used for measuring the score of each query by metrics described in subsection 2.3. Each section has about 2000 training examples and 400 testing examples.

3.2 Baselines

Random Selection (RS) - We randomly select 5 queries from the set of queries obtained after the query reduction step. Then, we average their scores obtained using best query selection metric described in subsection 2.3.

Section Heading Query (SHQ) - In this baseline method, we average the score of queries obtained using section heading only, like "Brad Pitt career". This approach of formulating queries has been used in [1] with an assertion that it yields better performance than the queries extracted from the body text.

Query Patterns (QP) – In this baseline method, we average the score of the query patterns extracted by the proposed approach of Tanaka et al. [2].

We use the same scoring metric as described in subsection 2.3 for each of the baselines.

3.3 Results

Table 1 shows the scores of our proposed method and the baselines. Our method outperforms the baselines in each section of the template for each category. Our proposed method extracts queries from the body text and beats the scores of queries formed using section heading by a comfortable margin. Although, our proposed approach performs better than [2], the baseline QP still gives good scores than rest of the baselines. It shows that queries extracted from the body text can retrieve more relevant information than those extracted from the section titles. The best

Table 1. Performance of the queries for the category -

a) American Male Actors

Sections	RS	SHQ	QP	Proposed Method
Personal Life	0.35	0.33	0.36	**0.48**
Career	0.44	0.44	0.48	**0.50**
Early Life	0.47	0.46	0.49	**0.54**
Television	0.35	0.30	0.38	**0.48**

b) City and Towns in India

Sections	RS	SHQ	QP	Proposed Method
Geography	0.41	0.31	0.45	**0.48**
Demographics	0.41	0.50	0.50	**0.52**
Education	0.29	0.28	0.35	**0.44**
Transport	0.32	0.25	0.35	**0.46**

Table 2. Final queries selected for each section of the category Cities and Towns in India

Geography	"average elevation", "town city", "town area"
Demographics	"average literacy", "total population", "census total", "males females"
Education	"school colleges", "college institute", "engineering medical", "university institute"
Transport	"connected road", "city railway", "railway station", "airport station", "major city"

queries selected are very diverse and section-specific as shown in Table 2.

4. CONCLUSION

We presented a novel method which is simple yet effective to obtain a set of Web queries for retrieving relevant information regarding an entity. The experimental results show that the proposed method outperforms the baselines and thus, performs well for generation of queries from the body text. A text summarization application can be built based on the proposed method to ensure that the selected queries can generate a comprehensive summary for an entity given its Wikipedia category.

REFERENCES

[1] Sauper, C., and Barzilay, R. 2009. Automatically Generating Wikipedia Articles: A Structure-aware Approach. In Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Intl Joint Conf. on NLP of the AFNLP: Volume 1 - Volume 1, pages 208-216

[2] Tanaka, S., Okazaki, N., and Ishizuka, M. 2010. Learning Web Query Patterns for Imitating Wikipedia Articles. In Proc. of 23rd Intl Conf. On Computational Linguistics (COLING 2010) – Poster Volume, pp. 1229-1237.

[3] Sutskever, I., Martens, J., and Hinton, G. Generating Text with Recurrent Neural Networks. ICML 2011

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013