# 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

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 (<p></p>) tags in these webpages. A query's score is defined as:

$$\text{query score, } q = \qquad\qquad (1)$$

Where "*m*" is the total number of articles in testing dataset and "  " 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.

$$= \mathit{max(cosine\_sim(i,\ section\_body_j\,))}, \qquad\qquad (2)$$

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 those l3(s)]os0564Tf( 3.52 hs0dposand sectf1-specific607(as shown6ind[(each)ose 11(s)]ose 22 e s- 3.52 rns15.3 dposTable 2)an.ion each)/TT0 1 Tf8ion