Information Retrieval and Extraction

Final Semester Examination

18th November 2019

Three Sections **180 Minutes** **180 Marks**

Section I: (Answer ALL FOUR questions – 60 Marks)

1. What is the difference between Taxonomy, Ontology and Knowledge Graph (KG)? Give examples and explain three generic and three vertical KGs. Is it possible to combine multiple KGs to create a larger KG? If so, what are the challenges?

2. Define the scope of the problem and propose the solution and evaluation criteria for "Entity disambiguation". E.g. Mahatma, Bapu, Gandhi, Gandhiji, M.K.Gandhi all are referring to the same entity in the knowledge base.

3. If you want to develop a solution for fine grained classification task such as sexism classification discussed in the class, using deep learning, what would be your proposed architecture?

4. How RNNs are used for language modeling? What are the differences between n-gram models and RNN based models in generating natural language text?

Section II: Major Project – Peer learning (60 Marks) – answer any two questions

Please address three perspectives in your answer:

- *Scope of the problem – challenges*
- *Solution as given by your peer group(s) – your understanding in depth is expected. If two groups have solved it, give both solutions*
- *Comparison, analysis and your own evaluation of the solution proposed wrt the problem*

1. Rumour detection

Description: Task is to label the type of interaction between a given statement (rumourous post) and a reply post (the latter can be either direct or nested replies). Each tweet in the tree-structured thread will have to be categorised into one of the following four categories: Support, Deny, Query, Comment.

2. Identifying topics/pages to be included in Indian Language Wikipedia

Description:
Use Indian language newspapers, text books and current affairs content as a set of sources to identify the new topics to be added to any Indian language Wikipedia. In order to qualify to be a Wikipedia page, there has to be enough evidence or set of references - what are the topics that has these kinds of evidences/references but does not have a Wikipedia page

3. CL-AFF Shared Task: In pursuit of happiness

Description: "GIVEN: An account of a happy moment, marked with individual's demographics, recollection time and relevant labels.

TASK: WHAT ARE THE INGREDIENTS FOR HAPPINESS ?

Semi-supervised learning task: Predict agency and social labels for happy moments, based on a small labelled and large unlabelled training data.

Optional: HOW CAN WE MODEL HAPPINESS?
Unsupervised task: Propose new characterizations and insights (not necessarily and not limited to themes) for happy moments in the test set, e.g., in terms of affect, emotion, participants and content.

Section III Research Paper (60 Marks)

Read the attached short research paper on "Towards Semantic Retrieval of Hash tags in Microblogs" and answer the following questions:

1. Write a 500-word summary of the main contributions of the paper in terms of problem definition, solution outline, and evaluation.

2. What are the IR/IE specific contributions?

3. What are the Machine Learning specific contributions?

4. What are the limitations of the approach?

5. What improvements can be made the evaluation described in the paper?

6. What are the possible uses/applications of this work? Try to list at least five with short descriptions.

Towards Semantic Retrieval of Hashtags in Microblogs

Piyush Bansal Somay Jain Vasudeva Varma
International Institute of Information Technology
Hyderabad, India
{piyush.bansal@research, somay.jain@students}.iiit.ac.in, vv@iiit.ac.in

ABSTRACT

On various microblogging platforms like Twitter, the users post short text messages ranging from news and information to thoughts and daily chatter. These messages often contain keywords called *Hashtags*, which are semantico-syntactic constructs that enable topical classification of the microblog posts. In this poster, we propose and evaluate a novel method of semantic enrichment of microblogs for a particular type of entity search – retrieving a ranked list of the top-k hashtags relevant to a user's query Q. Such a list can help the users track posts of their general interest. We show that our technique significantly improved microblog retrieval as well. We tested our approach on the publicly available Stanford sentiment analysis tweet corpus. We observed an improvement of more than 10% in NDCG for microblog retrieval task, and around 11% in mean average precision for hashtag retrieval task.

1. INTRODUCTION

Microblogging services enable communication at a massive scale. Twitter, one of the most popular microblogging services reportedly has 284 monthly active users, who post around 500 million short text posts (limited to 140 characters) called "tweets" everyday. Often users include keywords prefixed with '#', called *hashtags* to indicate and organize the contextual meanings of their tweets. Hence, to retrieve information related to a user's interest, for instance, "Rock concerts", it'd be very helpful to the user if they can be suggested a list of hashtags which are commonly used in relation to "Rock concerts". By tracking these hashtags, a user can gain information about rock concerts via the posted tweets. However, it's not possible for the user to manually figure out all the hashtags that are used across Twitter, relevant to their interest. In this poster, we address this problem and present a system (publicly accessible at `http://bit.ly/SemanticHashtagRetrieval`) that takes a query Q from the user and returns a ranked list of hashtags most relevant to Q.

In the past, microblog retrieval has been studied in various contexts: Sakaki et al. [5] exploit the real-time nature of tweets to discover events. TweetMotif [4] presents an exploratory search interface to deal with microblog retrieval, trying to summarize topics by analyzing co-occurrence patterns. More recently, Efron et al. [3] have studied hashtag retrieval from a query expansion point of view. However, almost all these retrieval approaches are strictly term-based, which are sensitive to polysemy, and term-use variation. In this poster, we propose "Semantically Enriched Microblog Document (SEMD)" structure, which enables semantic retrieval of hashtags and microblog posts, trying to overcome these limitations to a greater extent. Table 1 shows the top two hashtags retrieved for a small subset of queries.

Query	Baseline	Our Approach
Rock Music	#music #rock	#hardrock #progrock
TV Shows	#tv #chuck	#tvseries #addictivetvshows
Space Research	#space #research	#nasa #space

Table 1: Retreived hashtags for various queries

Tweet Text Excluding Hashtags (f_{TT}): **string**	
Hashtags(f_H): **keyword**	Segmented Hashtags(f_{SH}): **keyword**
Lead paragraph of Wikipedia linked entities in Segmented Hashtags (f_{LSH}): **string**	
Lead paragraph of Wikipedia-linked entities in Tweet Text (f_{LTT}): **string**	

Table 2: Semantically Enriched Microblog Document (SEMD) Structure

2. SEMANTIC ENRICHMENT

Traditionally most of the microblog IR has either ignored hashtags for analysis, or has treated them as single words. However, that might not be true for most of the hashtags - *#WWW2015Firenze* refers to "WWW 2015 Firenze". Recent work by Bansal et al. [1] presents a machine learning based approach to segment the hashtags and link the entities in hashtags to Wikipedia. This allows to extract latent semantic information about hashtags. It's worth noting here that their proposed approach also performs entity linking on the rest of the tweet text. We follow a similar approach with a few modifications to reduce latency and ensure high throughput, which is critical for a real time search engine such as Twitter. Notably, we make the following modifications –

1. Unlike Bansal et al. [1], we replace Microsoft Web N-

WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742717.

Gram Services[1] to compute n-gram word probabilities with in-memory Yahoo! Webscope N-gram dataset[2].

2. We pick only top three features - Unigram score, Bigram score, and Relatedness score, out of the originally proposed five features in [1]. This is motivated by the feature contribution results as reported in [1]. It's interesting to note that the features that contribute the most are less computation intensive.

These modifications help us to build a system capable of extracting latent semantic information from hashtags as well as the tweet text.

Semantically Enriched Microblog Document: We propose a virtual document structure that is enriched with semantic information obtained as described in the above section. The proposed document has 5 fields as can be seen in Table 2. We use Whoosh[3] library's BM25F implementation to retrieve microblogs most relevant to a given query Q. BM25F associates weight to each document field inversely proportional to it's length. Hence, shorter fields like f_{TT}, f_H and f_{SH} have more weight in retrieval process. Our premise is that these fields have greater importance than f_{LTT} and f_{LSH}, since hashtags tend to be the gist of the tweet.

3. RETRIEVAL PROCEDURE

For a given user query Q, we obtain a list of top 500 microblogs ranked by their relevance according to SEMD structure. In order to retrieve most relevant hashtags, we propose and experiment with a few hashtag ranking approaches -

1. *GlobalRank (GR)*, where the hashtags in top 500 retrieved microblogs were ranked on the basis of their frequency in the overall corpus.
2. *RetrievedHashtagRank (RHR)*, where the hashtags in top 500 retrieved microblogs were ranked on the basis of their frequency in top 500 microblogs.
3. *TF-IDFRank (TFIDFR)*, where we iteratively refine the quality of hashtags retrieved, while boosting the recall. We used TF-IDF weights for pseudo relevance feedback.
4. *KLDivergenceRank (KLDR)*, where we use KL Divergence for blind feedback.

We observed that KLDR performed significantly better than the other proposed approaches. A detailed comparative analysis is present in Table 3.

4. EXPERIMENTS AND RESULTS

Dataset: We used Stanford Sentiment Analysis tweet corpus[4] for testing our approach. The dataset consists of 1.6 million tweets collected between April 6, 2009 and June 25, 2009.

Experiments: We perform our evaluation for two tasks - 1. Hashtag Retrieval, and 2. Microblog Retrieval. 5 experienced Twitter users were explained the problem statements. Subsequently, they were asked to suggest 10 queries each, relevant to the given tasks, hence obtaining a list of total 50 queries. We define the baseline as BM25F search conducted on original microblog posts. The users were asked to search for the 50 queries on our search system, ranking each result on a 5 point Likert scale. Our evaluation interface presented the results from baseline, as well as from the proposed SEMD search engine. The search results were anonymized, and randomized for each query so as to prevent cognitive bias caused by learning effects between multiple queries. For both the tasks, we presented the user with top 10 results relevant to the user query Q.

[1] http://weblm.research.microsoft.com
[2] http://webscope.sandbox.yahoo.com
[3] https://pypi.python.org/pypi/Whoosh
[4] http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

Ranking Model	Baseline			Our Approach		
	MAP	MRR	NDCG	MAP	MRR	NDCG
GR	0.392	0.557	0.55	0.451	0.686	0.591
RHR	0.435	0.577	0.572	0.499	0.698	0.670
TFIDFR	0.475	0.687	0.691	0.569	0.713	0.750
KLDR	0.567	0.689	0.794	0.674	0.759	0.831

Table 3: Comparative results for hashtag retrieval

	Rel. Recall	MAP	MRR	NDCG
Baseline	NA	0.761	0.850	0.797
Our Approach	0.845	0.815	0.886	0.898

Table 4: Comparative results for microblog retrieval

Results and Analysis: We report Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) for both the tasks in Tables 3 and 4 respectively. Our definition of these metrics is same as proposed by Croft et al. [2]. MRR doesn't vary much between the different retrieval approaches, since an acceptable result (Likert scale score >=4) was observed in the first few results for almost all the retrieval approaches. However, other measures like NDCG, and MAP show encouraging improvement over the baseline and [3]. It's important to note here that these improvements were observed to be statistically significant ($p < 0.05$).

5. CONCLUSIONS AND FUTURE WORK

We have presented and evaluated a semantic search system in context of hashtag and microblog retrieval. We demonstrate how our approach is better than the existing approaches by detailed experiments. In the future, we'd experiment by enriching microblog posts with additional semantic information. This would include data mined from external links present in the microblog posts, author information and location et cetera.

6. REFERENCES

[1] P. Bansal, R. Bansal, and V. Varma. Towards deep semantic analysis of hashtags. In *Advances in Information Retrieval*. Springer, 2015.

[2] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, 2009.

[3] M. Efron. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[4] B. O'Connor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter, 2010.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.