Comparing N-gram Models for Tag Suggestion in Tagging System

*Hyunwoo Kim, *Kangpyo Lee
Computer Science & Engineering
Seoul National University
Seoul, Korea
*{hwkim, kplee}@idb.snu.ac.kr

Hyopil Shin
Department of Linguistics
Seoul National University
Seoul, Korea
hpshin@snu.ac.kr

Hyoung-Joo Kim
Computer Science & Engineering
Seoul National University
Seoul, Korea
hjk@snu.ac.kr

***Abstract*— On the web, a *tag* is a significant keyword of content, including a photo, video, and blog article. A *tagging* is an action of adding a tag to content. Many web sites, such as del.icio.us [1] and CiteULike [2], are providing tagging system. These tags can be used for web search. Users have already recognized the value and the importance of tags, but some users do not use tags. These users might feel annoyed to be forced to add tags, or they might simply not know what to add in order to obtain a good search result. These problems are the reasons why tag suggestion system would be beneficial. In this paper, we use n-gram models for tag suggestion in tagging system. We gathered tag data from various web sites. Based on crawled tag data, we will employ various n-gram models and compare obtained results in the next paper. This is a progress paper.**

Keywords- Web 2.0; Folksonomy; Tag Suggestion; Tagging; N-gram; Natural Language Processing

I. INTRODUCTION

Users on the web create thousands of multimedia content every day. They upload their videos or photos to web sites, such as YouTube and Filckr. Because there are too much content on the web, people have to search when they want to find some content on the web. In articles or blog posts, search engine could gather information from a text itself. However, there is no information contained in multimedia content except a title. The title might not be able to include all information. Given this situation, a tag, a meaningful keyword which describes corresponding content, would be beneficial for web search.

The tag is efficient index for web search. It could describe information of content, even if the content has no text information. If there is a picture of a sea titled 'Blue Ocean' on the web, when some users want to know the location of the ocean, the user could not obtain the exact location because there is no information except the title 'Blue Ocean'. If there are tags such as *Hawaii*, *blue sea*, and *beach* for 'Blue Ocean' file, these tags would be useful information to find the location of the picture. Tags could be valuable to search appropriate results in tagging system.

In spite of the value and the importance of the tag, the number of users who use tags is relatively of small. Some people use tags for their content, but others do not use tags. These users might not be familiar with tags, or they might not be aware of tags. Because of these problems, tag suggestion system is needed.

.

II. COMPARING N-GRAM MODELS

In this section, we explain n-gram models and propose n-gram models for tag suggestion method in tagging system.

A. N-gram Models

N-gram models are widely used in statistical natural language processing [1]. An n-gram is a substring of n characters or a substring of n words in a given text. For instance, a bigram is 2-word substring of a text and a trigram is 3-word substring. If there is a sentence 'Thank you very much', bigrams are {Thank, you}, {you, very}, and {very, much}. Trigrams are {Thank, you, very} and {you, very, much}. When a sentence consists of *m* words, it is possible to make *m*-gram at most. N-gram models are used in indexing method. Using trigram, a word 'thank' can be indexed by *tha*, *han*, and *ank*. N-gram models are also used for prediction. This method calculates probabilities from corpus data. Given n-1 words, n-gram models predict which word would be next word based on calculated probabilities. For example, n-gram models predict next word of given 'thank you very' sentence based on previous bigram or trigram examples. It could be 'much'.

B. Tag Suggestion

Some web sites, including del.icio.us, provide tag suggestion in their tagging system. Tagging system in del.icio.us is a collective tagging system. In collective tagging system, more than 2 users can add tags to same

[1] http://www.delicious.com
[2] http://www.citeulike.org

content. In del.icio.us, tag suggestion method suggests tags which are already added by other users based on the popularity. Most tag suggestion methods are based on tag co-occurrence [2, 3]. These methods count co-occurrence frequency between two tags and calculate correlation between two tags. In these methods, all tags in a tag set are regarded as co-occurred tags. However, n-gram models consider not all tags but adjacent tags as co-occurred tags.

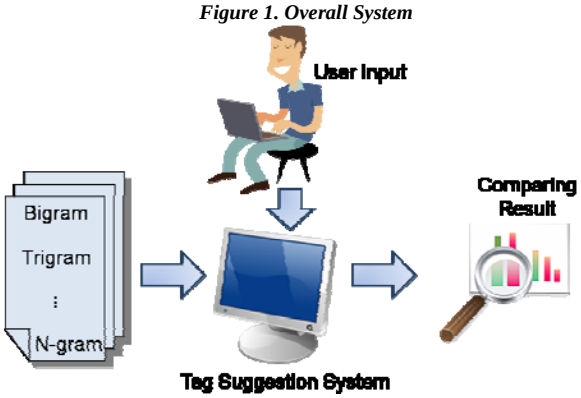
C. *Process of Thinking*

When people express their thought, they use association of ideas. They tend to think next idea connected with previous idea. We assume that this process would happen when people add tags to content. Because the tags are the descriptor of the content and these tags are from user's thoughts, adjacent tags are more related than other tags in a series of tags. In a tag set, for example, {*Apple*, *iPod*, *iTunes*, *music}*, *Apple* and *iPod* has closer connections than *Apple* and *music*.

III. OUR APPROACH

Current tag suggestion research studies are focusing on tag co-occurrence. People develop their thought using association of ideas. Correlation between adjacent tags would be much higher than other tags. Given this situation, various n-gram models, including bigram and trigram, is beneficial to find correlated tags for tag suggestion because these models consider neighborhood tags as co-occurred tags.

Figure 1. Overall System



Our tag suggestion approach calculates tag co-occurrence count using various n-gram models from bigram to *n*-gram. Bigram takes count of tag co-occurrence between adjacent tags and trigram takes count of three tags at a time. If a tag set consists of *n* tags, *n*-gram would consider all tags as co-occurred tags.

After calculating tag co-occurrence count, our tag suggestion system stores the date in database. When a user adds tags to content, our system recognizes the context of users input. If the user adds one tag, our system finds the most co-occurred tags with the user entered tag in the database. The result of bigram model would be different from that of trigram model. To evaluate our method, we provide training set and test set. The training set is used for calculating tag co-occurrence count. The test set is divided into user input and answer set. For tag suggestion, our suggestion system employs both user input of the test set and calculated co-occurrence count. Suggested tags are evaluated by the answer set of the test set. We will compare the results of various n-gram models.

Each n-gram models has not only strength but also weakness. Bigram would capture the most related tags but the result could be rare. Trigram would gather more results than bigram but the correlation between suggested tags could be weaker than bigram. The number of suggested tags in *n*-gram would be large because this model considers all tags as co-occurred tags. However, the relation among suggested tags could be weakest than any other method.

There would be the most efficient n-gram model. It could be bigram or it could be 5-gram. We will figure out which n-gram model is superior to other models by analyzing experiment results in the next paper.

ACKNOWLEDGMENT

This research was supported by the Brain Korea 21 Project, the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement (grant number IITA-2008-C1090-0801-0031), and a grant (07High Tech A01) from High tech Urban Development Program funded by Ministry of Land, Transportation and Mari-time Affairs of Korean government.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing*: Pearson Education International, 2008.

[2] N. Garg and I. Weber, "Personalized tag suggestion for flickr," in *WWW*, 2008.

[3] B. Sigurbjornsson and R. Zwol, "Flickr Tag Recommendation based on Collective Knowledge," in *WWW*, 2008.