Junfang JIA, Guoqiang LI, 2021. Learning natural ordering of tags in domain-specific Q&A sites. *Frontiers of Information Technology & Electronic Engineering*, 22(2):170-184. https://doi.org/10.1631/FITEE.1900645

Learning natural ordering of tags in domain-specific Q&A sites

Key words: Question and answering (Q&A) sites; Tagging; Natural order; Skip gram

Corresponding author: Guoqiang LI
E-mail: li.g@sjtu.edu.cn
ORCID: Junfang JIA, https://orcid.org/0000-0002-3451-8487; Guoqiang LI, https://orcid.org/0000-0001-9005-7112

Motivation

· Tagging is essential to users of social computing systems (e.g., Q&A sites), and existing work has shown that users can develop implicit consensus about the choice of tags.

· However, there has been no work studying the regularities in how users order tags during tagging, which could be helpful in improving existing tag recommendation and Q&A site navigation.

Main idea

We hypothesize that:

· the same argument applies to tags used to describe questions in a domain-specific Q&A site;

· “tag sentences” that Q&A site users use to describe questions are mostly simple and rather repetitive;

· such “tag sentences” have predictable ordering that can be captured in statistical language models.

Main idea (Cont'd)

We validate our hypothesis through the following three research questions:

· Is there any natural ordering of tags in CodeProject?

· Why natural ordering of tags (if any) emerges from users' tagging activities?

· Is the natural ordering of tags (if any) domain-specific or domain-agnostic?

Method for Q1

Is there any natural ordering of tags in CodeProject?

· Using the corpus of tag sequences, we train a *K*-skip bi-gram model and perform 10-fold cross-validation of the model for tag sequence prediction.

· As no models can predict a totally random distribution, if the prediction of a tag sequence matches the original tag sentences in most cases, we can infer that some natural ordering of tags exists.

Major results for Q1

· When the *K*-skip is more than three, 94% of the predicted tag sequences are exactly the same as the original sequence.

· Thus, we conclude that natural ordering exists in tag sentences in CodeProject.

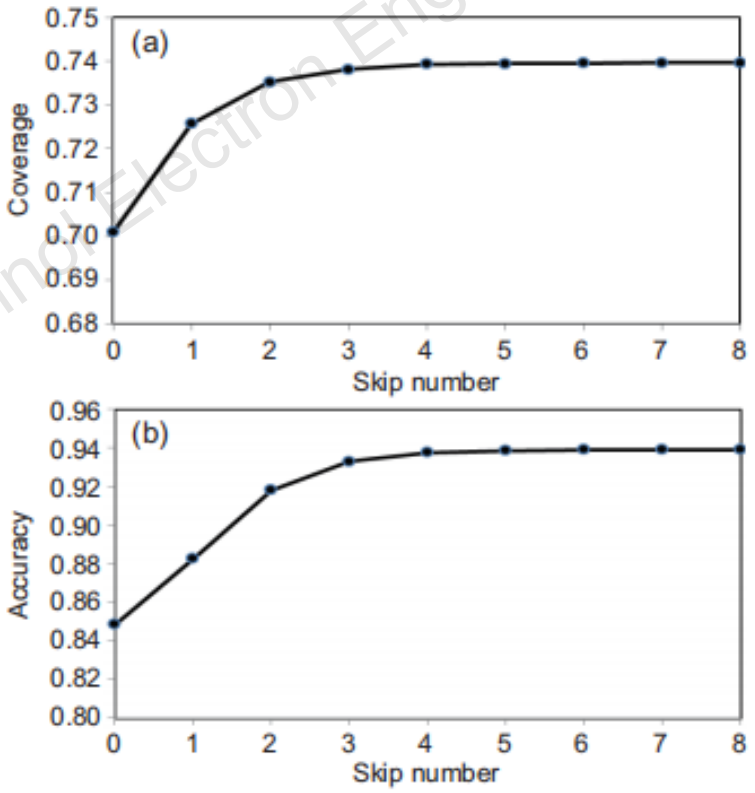

Fig. 3 Coverage (a) and accuracy (Levenshtein distance=0) (b) with different skip numbers

Method for Q2

Why can natural ordering of tags emerge from users' tagging activities?

· We identify the underlying factors that result in such natural ordering of tags by examining the relationships of tags in all the bi-grams in the 3-skip bi-gram model.

· For each pair of bi-gram and its corresponding order-reversed bi-gram, we compute a frequency quotient *q* by dividing the frequency of the less frequently occurring bi-gram by the frequency of the more frequently occurring bi-gram.

Major results for Q2

· As low-quotient bi-grams account for 96.5% of all bi-grams, the presence of inclusion and convention explains why natural ordering of tags can emerge from user tagging behavior.

· Most of our prediction errors are indeed caused by such juxtaposed bi-grams.

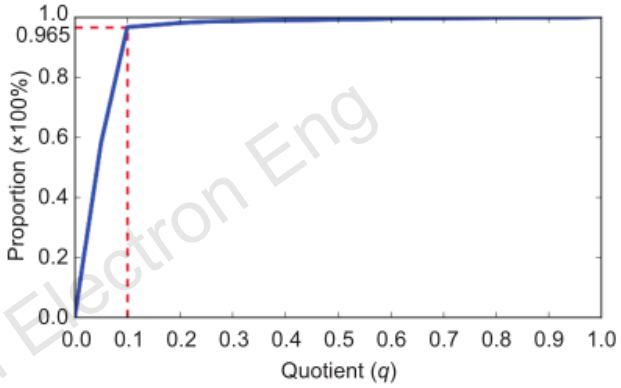

Fig. 5 Cumulative distribution function of frequency quotient of bi-grams, where the vertical axis represents the proportion of the bi-grams taking on a quotient less than or equal to q. The dashed line shows that quotient of 96.5% of bi-grams is lower than 0.1

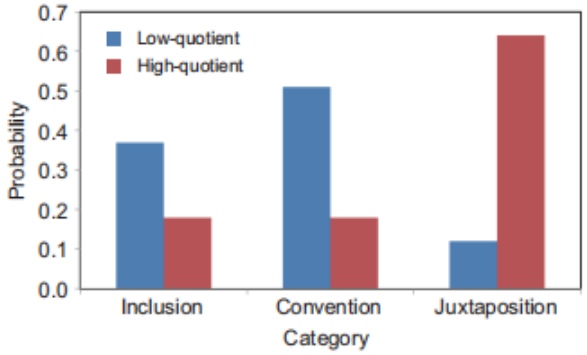

Fig. 6 Proportion of different categories of relationships of the sampled 400 low-quotient bi-grams and 100 high-quotient bi-grams (References to color refer to the online version of this figure)

Method for Q3

Is the natural ordering of tags domain-specific or domain-agnostic?

· We perform the same 10-fold cross-validation of the *K*-skip bi-gram language model on the tag sentences extracted from CareerCup (interviews), Biostars (bioinformatics), and SegmentFault (Chinese programming) questions.

· Next, we examine the relationships of tags in bi-grams in the language model of CareerCup, Biostars, and SegmentFault.

Major results for Q3

· On one hand, order exists in tags of the other three Q&A sites.

· On the other hand, our language model is a feasible way to capture such implicit ordering.

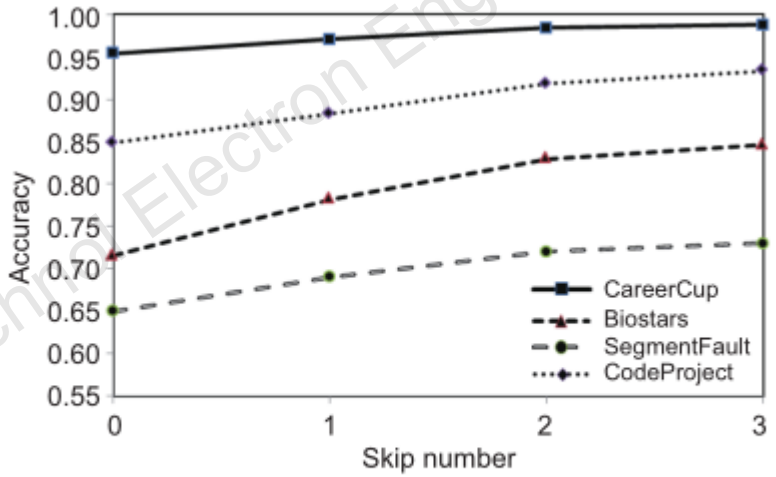

Fig. 9 Accuracy (Levenshtein distance=0) under different skip numbers

Conclusions

· Tag sequences of large collections of questions in domain-specific Q&A sites have implicit natural ordering.

· Inclusion and convention usually result in one ordering of tags being preferred by the community over alternative ordering, while juxtaposition usually results in an arbitrary ordering of tags.

· The language model of tag sequences can automatically reorder tags in a way that accurately matches the tag-reordering actions by high-reputation users in the community.

· Natural ordering of tags is largely domain-agnostic, but different Q&A sites may exhibit different ordering characteristics because of the nature of the domain.

Ms. Junfang JIA is an associate professor in School of Computer and Network Engineering, Shanxi Datong University. She received the BS and MS degrees from Shanxi University in 2000 and 2011, respectively. Her research interests include cluster analysis, machine learning, etc.

Dr. Guoqiang LI is an associate professor in School of Software, Shanghai Jiao Tong University. He received the BS, MS, and PhD degrees from Taiyuan University of Technology, Shanghai Jiao Tong University, and Japan Advanced Institute of Science and Technology in 2001, 2005, and 2008, respectively. He worked as a postdoctoral research fellow in Nagoya University, during 2008-2009, as an assistant professor in Shanghai Jiao Tong University, during 2009-2013, as an academic visitor in University of Oxford, during 2015-2016, and as a guest associate professor in Kyushu University, during 2016-2020. His research interests include formal verification, programming language theory, and knowledge representation and reasoning.