Improving Accuracy of Tagging Systems Using Tag Qualifiers and Tagraph Vocabulary System

Syavash Nobarany[1], Mona Haraty[1]

[1]University of Tehran
Electrical and Computer Engineering Department
{Nobarany, Haraty}@gmail.com

Abstract: This short paper addresses the lack of accuracy in tagging systems in comparison with traditional search-engines. Lack of accuracy is caused by the Vocabulary Problems and the nature of these systems which rely on lower number of index terms for each resource. Tagraph, a weighted directed graph of tags, and Tag Qualifiers are proposed to mitigate these problems. Both solutions are based on community contributions, therefore specific procedures such as task routing can be used to take the most advantage of the contributions.

Keywords: vocabulary, tag, Tagraph, Tag Qualifier, accuracy, Tagging system

1. Introduction

Collaborative tagging has become increasingly popular as a common way of organizing, sharing and discovering information. Properties and design of tagging systems are investigated in [1, 2, 3]. One of the important deficiencies of current tagging systems is the lack of accuracy about the way and the quantity with which a resource relates to a tag. Although in tagging systems keywords are detected with less error, the lack of accuracy in tagging leads to poor search results. In traditional search engines, relevance of a term to a resource is estimated using several methods such as calculating term frequency,. Tagging systems have no competitor in multimedia information retrieval. Solving accuracy problems of Tagging systems is one of their important challenges to become a robust competitor for traditional search engines in text retrieval.

Another class of problems that are not limited to tagging systems and involve all information retrieval systems is the Vocabulary Problems, introduced by George Furnas et al [7]. An example of the Vocabulary Problems happens where different users use different tags to describe the same resources, which can lead to missed information or inefficient user interactions. We introduce a community-based approach to overcome the Vocabulary Problems without requiring either the rigidity and steep learning curve of controlled vocabularies, or the computational complexity of automatic approaches to term disambiguation.

2. Tag Qualifiers

Tags have several properties, which we call Tag Qualifiers. A basic qualifier, which can improve search results, is the Relevance Qualifier. Tags that are assigned to a resource are not equally relevant to the resource. Therefore, Relevance Qualifier can not be defined to express their relevance. Relevance Qualifier is a key element for improving *precision* of tagging systems, which is defined as the number of relevant retrieved documents divided by the number of all retrieved documents.

Another qualifier, which is already being used in tagging systems, is the Restriction Qualifier which may be implemented using a public-private property or, in a more precise manner, by mapping users to sets of permissions.

Beside general tag qualifiers, context-specific qualifiers can also be helpful in special purpose tagging systems. For example, in virtual learning environments, level of educational materials can be considered as a qualifier. Bookmarks, which are used to express user interest in a resource, also may have qualifiers. For example, Recommendation Qualifier can be used to express users' interest in a resource. This qualifier leads to a better ranking of results. Choosing the appropriate set of qualifiers can be considered as a part of the design of tagging systems.

2.1 Auto-assigned Relevance Qualifiers

Assigning Relevance Qualifier is not easy for taggers. Most users are not comfortable with using tag-qualifiers except when they are using them for special purposes such as using restriction qualifiers. More detail on incentives in online social interaction can be found in [5] and [6].

A simple yet robust solution to this problem is using auto-assigned relevance qualifiers that can be calculated automatically using traditional information retrieval methods such as calculating term frequency usage and document frequency retrieval. This will reduce user obligation to assign Relevance qualifiers.

3. Tagraph: Graph based vocabulary system

A common assumption in many information retrieval systems such as probabilistic systems is considering terms' occurrences independent from each other, which is incorrect in many situations. A famous counterexample for this assumption occurs when different users use different tags to describe the same resources. This is known as "synonyms problem". However, this problem is not limited to synonyms. Consider a document tagged with the term 'EJB', and a query contains the term 'Java'. In this situation, EJB is not a synonym of Java, but their meanings are related. The document tagged with Java, however, may not necessarily be related to EJB. To deal with situations similar to the above example, Tagraph, a graph with tags as nodes and relationships between tags as arcs, is proposed. In Tagraph, arcs are directed since the relationships between tags are not necessarily two-way, as in the previous example.

Tagraph is a weighted graph based on the strength of the relationship between a tag and its implementation. For example the relationship between JSF and MyFaces, an implementation of JSF, is stronger than the relationship between Java and JSF. Relevance arcs can increase the *recall* of tagging systems, which is defined as the number of relevant retrieved documents divided by the number of relevant documents.

Relevance arcs are transitive and their weights can be used to find out the relevance of two nodes that are not directly connected to each other. Transitive relevance can be inferred using Equation 1. In this formula, source of the non-existent relevance arc is denoted as *s*, its destination is denoted as *d* and tags that have direct relevance arc to source are denoted as *i*. This equation finds the strongest relationship between source and destination.

$$r_{sd} = MAX_i(r_{si} \times r_{id}) \qquad \textbf{(1)}$$

Homonyms problem is another important problem that influences tagging systems more than traditional search engines. This problem stems from the smaller number of index terms that are used to describe each document in tagging systems. Large number of index terms in traditional search engines can determine the context of a document, while in a tagging system many documents have less than three tags. Therefore, defining context of the tags deems necessary in tagging systems.

To overcome this problem, each node of the Tagraph has a set of context-definition terms. These terms are chosen from a predetermined hierarchy of terms. This solution can increase *precision* of tagging systems. A sample Tagraph is depicted in Figure 1.

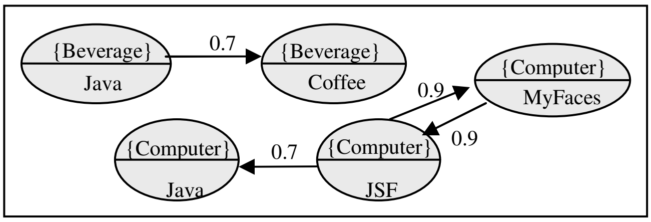
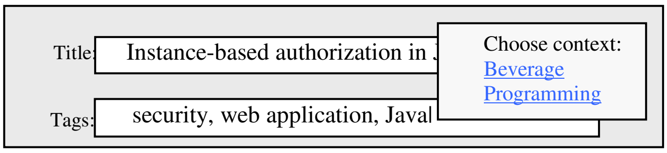


Fig. 1. Sample graphs in Tagraph system

Proposed system's performance is highly dependent on user interface design. Figure 1. shows a UI prototype, which allows context definition while tagging a resource. Similar procedure can be used in the query formulation UI.



Title: Instance-based authorization in J

Tags: security, web application, Java|

Choose context:
Beverage
Programming

Fig. 2. AJAX based tagging UI with context selection

3.1 Discovering Tagraph edges and their weights

Finding and assigning relevance edges can be done automatically using WordNet. A community-based approach is needed to compensate for WordNet deficiencies such as inconsistencies. A system like SuggestBot introduced by Cosley et al. [4] which has increased the number of contributions in Wikipedia, can be used to take the most advantage of taggers. Task definition for Tagraph requires automatic document inspection to derive possible edges from co-occurrence of tags. Determining possible edges, assigning their weight can be suggested to taggers based on their history.

4. Conclusions

In this paper, we suggest using Tag Qualifiers to improve precision of tagging and search results through storing more information about each tag. Auto indexing also can be used to help taggers to assign relevance-qualifier to tags.

The Tagraph vocabulary system is proposed to overcome the Vocabulary Problems of tagging systems. These problems are mitigated using a community-based approach and the weighted directed graph of tags and their relevance. Context declaring facilities are used to overcome homonyms problem. These solutions can improve precision and recall of tagging systems and a comprehensive evaluation is needed to verify the effectiveness of each solution.

Acknowledgements

We would like to thank Farid Vakil, Sam Vakil & Dan Cosley for their comments.

References

[1] Sen S., Shyong, Lam, T.K., Cosley, D., Rashid, A.M., Frankowski, D., Harper, F., Osterhouse, J., Riedl, J.: tagging, community, vocabulary, evolution. 20th anniversary conference on Computer supported cooperative work , pp. 181—190. ACM Press (2006)

[2] Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead., seventeenth ACM conference on Hypertext and hypermedia , pp. 31—40. ACM Press (2006)

[3] Golber, S., Huberman, B.A., The Structure of Collaborative Tagging System, Information Dynamics Lab: HP Labs, Palo Alto, USA, http://arxiv.org/abs/cs.DL/0508082

[4] Cosley, D., Frankowski, D., Terveen, L., and Riedl, J.:SuggestBot: using intelligent task routing to help people find work in wikipedia. 12th international Conference on intelligent User interfaces, pp. 181—190. ACM Press (2007)

[5] Korfiatis, N., Social and Economic Incentives in Online Social Interactions: A Model and Typology. 30th Information Systems Research Seminar in Scandinavia IRIS, (2007)

[6] Ludford, P., Cosley, D., Frankowski, D., & Terveen, L. (2004). Think Different: Increasing Online Community Participation Using Uniqueness and Group Dissimilarity. CHI 2004, pp. 631--638. ACM Press (2004)

[7] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. The vocabulary problem in human-system communication. Communications, 30 (11), 964 – 971. ACM Press (1987).