SemSearch 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online at CEUR-WS.org/Vol-334/

Resolving Lexical Ambiguities in Folksonomy Based Search Systems through Common Sense and Personalization

Mohammad Nauman[1], Shahbaz Khan[2], Muhammad Amin[3], and Fida Hussain[4]

[1] recluze@gmail.com
[2] shazalive@gmail.com
[3] clickforamin@gmail.com
Research Group Security Engineering
Institute of Management Sciences.
[4] fidamsse@gmail.com
City University of Science and Information Technology
Peshawar, Pakistan.

Abstract. Information on Web2.0, generated by users of web based services, is both difficult to organize and organic in nature. Content categorization and search in such situation offers challenging scenarios. The primary means of content categorization in such social services is folksonomy or collaborative tagging. During search in folksonomy, several issues arise due to lexical ambiguities in the way users choose tags to represent content. These are issues of different words representing the same concept, same words representing different concepts and variances in level of expertise of users. Past techniques to address these issues have worked on lexical analysis of term and have thus had only moderate levels of success. We have developed a model in which machine common sense and personalization is used to address these issues. In this paper, we explain our approach in detail, describe a prototype developed for the purpose of demonstrating feasibility of our approach and discuss an effectiveness study conducted to measure the success of our model. The results of the study are analyzed and future directions along this path of research are presented.

Key words: Common Sense, Folksonomy, Search, Web2.0.

1 Introduction

The social web is a collection of services providing user-created content. These are, among others, photo-sharing systems, blogs, wikis and image and map annotation systems. This collection of services is informally termed as Web2.0. Lack of a central organization for this huge amount of information is a significant hurdle that makes searching through Web 2.0 services very difficult. [1]

Categorization in Web2.0 service is based upon tags (or keywords), which make up a user-created organization. This organization of content is termed as folksonomy or more formally collaborative tagging. Tags serve as keywords

attached to a unit of content for the purpose of organization. Due to the reason that users assign tags to content based on their own experience, skill and mental state, several types of ambiguities arise in the categorization. Content retrieval in Web2.0 becomes very difficult in such a situation and several very important pieces of content might not be recalled due to these ambiguities.

Our study focuses on searching techniques for Web2.0 content and addressing the issue of ambiguity in search results. We have proposed a mechanism through which machine common sense can be used to automatically disambiguate tags and return more results which would otherwise be missed by traditional search mechanisms. The second aspect of our model focuses on user personalization in collaboration with machine common sense to increase the relevance of search results based on an individual users' preferences. Unlike some past techniques, our model requires a minimum of effort on the user's part and is thus very effective for system offering services to non-technical users.

The paper is organized as follows: First we describe the problems of lexical ambiguities in folksonomy based systems in detail. Then we discuss some related and background work which is relevant to our proposed model. Section 4 begins with a discussion of our model, describes how machine common sense and personalization can be used for the purpose of disambiguation in folksonomy and describes our model comprehensively. In Section 6 we discuss the effectiveness study conducted. Section 7 includes the results of the study and our thoughts on these results. Finally , we provide a few directions which can be useful in extending our model in the future.

2 Problem Overview

Web 2.0 services deals with huge amount of ever-growing and changing content. These services primarily depend on folksonomy for organization and retrieval of content.

Folksonomy being a very flexible technique also poses some serious drawbacks. The major problem with tagging is that it employs "folk psychology" to textually represent concepts. This problem branches off into two categories, Polysemy (using same word for different concept) and Synonymy (using different words for same concept). These vague variations are encountered due to the difference in inference of different users according to mental constructs such as knowledge and beliefs. To put it simply, this can be the difference of understanding of two or more users and/or different level of understanding of one user at different times. For example a picture of a car's interior can be tagged as "car", "automobile", "steering" or "leather". These problems arise while saving and retrieving of content.

Several strategies have been used to address the issues including those based on synonyms and co-occurrence frequencies. Since all these approaches are based on lexical analysis of terms instead of contextual, they have had only moderate levels of success [2].

Folksonomy is a non-hierarchical and non exclusive ontology. In such knowledge representation techniques, relationships between objects, concepts and other entities are fuzzy and boundaries between them are unclear.

Another problem with folksonomy (which it shares with traditional search systems) is that it does not provide other important sub-processes (facilities) in searching. The user has to examine the results, extract relevant information and take care of reflections and iterations during the search process.

Any search technique targeting folksonomy has to address all these issues. Traditional web search techniques, such as meta-search and/or categorization of contents into hierarchies, cannot be used because of flat ontological structure and loose textual representations. A more effective means of content retrieval might surface if certain non-traditional techniques are used. Our model uses a collaboration of two such techniques: machine common sense and personalization.

3 Related Work

Several techniques have been used for the purpose of solving issues of lexical ambiguities in folksonomy based services. The one closest to our approach of applying machine common sense was proposed in [3] and is called SemKey. It attaches semantics to tags associated with content. The tags are arranged in three relations: *hasAsTopcic, hasAsKind, myOpinionIs*. The user is expected to decide what attribute of the content they're tagging about. The SemKey system also disambiguates tags using WordNet when they're submitted. The issue with SemKey is that it expects users to associate more information with the content than just the tags. The beauty of folksonomy is that the users do not have to learn any formal mechanisms of content arrangement; instead, they can tag content using *freely chosen* words. We believe that whatever the mechanism for solving problems in collaborative tagging systems, this basic freedom should not be sacrificed. Instead, any technique used to address these issues ought to be automatic.

We have identified a technique developed by Liu et al. [4] which uses automated processes for personalization of search results. This basic technique uses search and access history for storing the user profile. The idea behind the approach is this: One user may associate a term, say "apple", with the category "cooking" while another may think of it as a brand. The user's profile and search history can be used to disambiguate the use of terms in such ambiguous cases.

Cat/Term	apple	recipe	pudding	football	soccer	fifa
COOKING	1	0.37	0.37	0	0	0
SOCCER	0	0	0	1	0.37	0.37

Table 1. Example representation of a user profile

User preference is maintained in a user profile matrix of weights, which consists of categories (representing the user's interest) and terms associated with these categories. A larger weight of a term for a category shows that the user normally associates the term with that category. We refer the reader to [4] for details regarding construction of this matrix.

4 Common Sense and Personalization for Folksonomy

Community generated tags are a vast source of information in a Web2.0 service. They are generated by users of the service and are heavily reflective of their own preferences, skill and common sense. This poses some serious problems for search in folksonomy.

We have developed a technique [5, 6] for applying machine common sense on search in folksonomy. The main strength of this technique is that it is based on contextual, not lexical, analysis of terms. The approach is based on query keyword expansion using a common sense knowledge base - the Open Mind Common Sense Project [7] - and a freely available common sense toolkit - ConceptNet[8].

The Open Mind Common Sense Project (OMCS) is a framework developed by Singh [9] for collecting common sense information from the general public using the world wide web as an interface. Since common sense is, by definition, bits of information shared by most people [8], it seems appropriate that everyone should be able to contribute to a common sense knowledge base. OMCS has had a lot of success over the years and has gathered more than 713,000 items of common sense information [10]. Several common sense reasoning tools [8, 11] have been extracted from the OMCS corpus among which ConceptNet [8] is the first. It is composed of more than 250,000 elements of common sense knowledge represented using natural language fragments and has 20 relation-types which include relations such as *PartOf*, *LocationOf*, *MotivationOf* etc. Two types of scores are assigned to each relation – f: number of times the relation occurs in OMCS corpus and i: number of times it was inferred from other fact.

Figure 1 shows an example [5] of concepts and links as used in ConceptNet.

We have identified the lack of contextual information and inference capabilities as the two major problems for search in folksonomy based systems. We believe that machine common sense can be used to address both of these problems. The basic common sense and folksonomy (CS&F) based search technique [5] works through concept expansion and a score function.

The technique expands concepts which are of a user-selected relation-type and have high conceptual similarity to user's search keyword. The value for conceptual similarity is given by:

$$C(x) = f(x) + (10 \cdot i(x)) \tag{1}$$

Search is performed for each expanded concept. Each *result item* may appear as a result item for more than one concepts (along with the associated search engine score S) and for each instance of this appearance, an instance score is calculated using a score function.

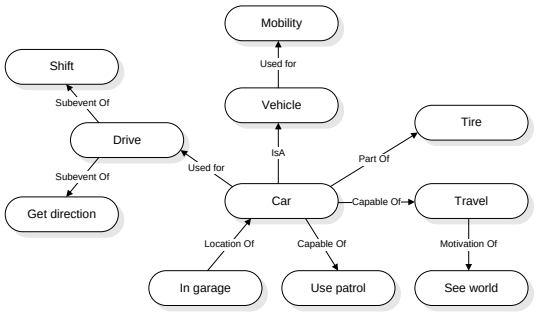

Fig. 1. Concepts related to CAR in ConceptNet

$$inst\_score(x_i) = (G \cdot \sigma(x_i)) + (1 - G) \cdot \gamma(x) \tag{2}$$

The total score of a result item is the sum of all instance scores:

$$score(x) = \sum_{i=1}^{n} inst\_score(x_i) \tag{3}$$

In this technique, two aspects are identified as leading to noise in search results:

- Polysemy: Very similar or even the same words may be used to define completely different concepts. Take for example the brand "Apple" and the fruit apple. Both of these concepts will be considered similar due to the shared lexical representation of the base concepts but for a user they are not similar.
- The score function is rudimentary and only assigns score based on generality and search engine score. Different users may find different results more relevant and therefore the results need some sort of personalization.

One method to address this issue is to use personalized web search for anticipating the user's categories of interest. The expanded concepts and ranked results can be tailored automatically for an individual user based on his/her search and access history. In a past work [6], we have studied this approach in detail.

5 Personalized CS&F Based Search

5.1 Concept Expansion

The personalized technique makes use of the category-term matrix M for concept expansion. Search and access history of a user can be used to personalize the results for individual users. There are two alternatives for using the search history

for concept expansion. One only expands concepts which are in the same category as the original keyword and the other assigns weights to all expanded concepts based on category similarity. The category (Φ_x) associated with a keyword x is that for which the column (T_x) representing the keyword has the highest value.

More precisely, let

Φ_o = Category of the original keyword
T_o = Column representing the original keyword
M_u = Matrix M for user u

then

Φ_o is that row for which

$$M_u(\Phi_o, T_o) = max(M_u(i, T_o)) \tag{4}$$

where i ranges over all rows of matrix M.

For concept expansion:

1. Calculate category for original keyword
2. Expand concepts through ConceptNet
3. Calculate categories for each expanded concept as in 4
4. For each category (k) (returned as result of Step 3), calculate category similarity (Θ) using the function:

$$\Theta_{e_k} = M_u(\Phi_o, T_{e_k}) \tag{5}$$

where

Φ_o is the category of the original keyword and
T_{e_k} is the column representing the concept e_k

5. Calculate *personalized conceptual similarity* by applying category similarity as a weight to the basic conceptual similarity given in 1.

$$C'(e_k) = C(e_k) \cdot \Theta_{e_k} \tag{6}$$

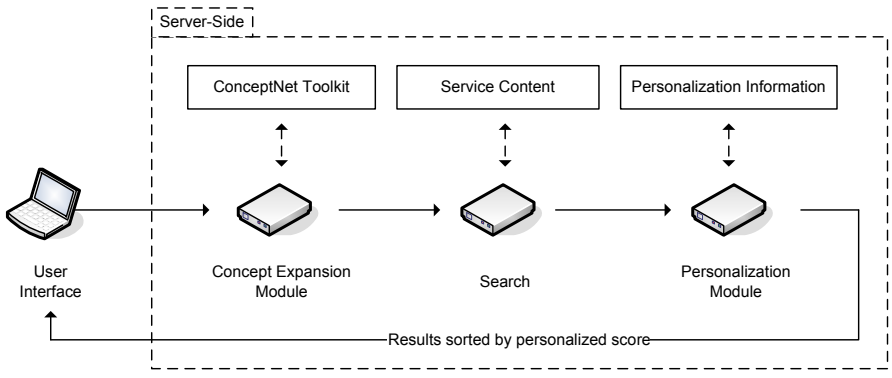

Fig. 2. Basic architecture of common sense and folksonomy based search systems [6]

6. Normalize the conceptual similarity – given as γ':

$$\gamma'(e_k) = \frac{C'(e_k)}{max(C'(e_k))} \tag{7}$$

5.2 Personalized Score Function

Once concepts are expanded, the score of the returned results can be recalculated to give *personalized score*. We note that there are usually more than one tags associated with a single piece of content. Personalized score is designed to take these different tags into account while ranking items. For each of these related tags, category similarity is calculated using the same function as in . We use r for *related* instead of e for *expanded*.

$$\Theta_{r_k}(x) = M_u(\Phi_o, T_{r_k}) \tag{8}$$

Finally, we *personalized score* ($score'$) is calculated as a function of the basic *score* and Θ_{r_k} given as:

$$score'(x) = \frac{score(x) + \sum_{k=1}^{n} \Theta_{r_k}(x)}{n+2} \tag{9}$$

Θ_{r_k} gives preference to those documents which are tagged with keywords belonging to the same category as the original search keyword. It also ensures that if a document is tagged with irrelevant keywords – say, the name of the user – the score is penalized.

5.3 Algorithm

Working of personalized web search in common sense and folksonomy based search systems is summarized in the algorithm described in Figure 3.

6 Effectiveness Study

A prototype of the proposed model showed the feasibility of constructing a search system based on the proposed model. To measure the effectiveness of the approach and the prototype, we conducted an effectiveness study.

The study aimed to gather quantitative results regarding the effectiveness of the search model. Since the intended audience of the system is the general public and not computer science experts, a questionnaire was developed which could be easily filled by non-experts and would provide us with quantitative results for drawing conclusions about the new technique. The sample size of the survey included 8 individuals from different levels of computer expertise. Data was collected through the use of a questionnaire hand-delivered to the participants. The questionnaires were filled by the participants while using the prototype and were returned in the same sitting. The important questions are given below along with their question numbers as given in the questionnaire:

Get search $keyword$ from user
Φ_o := getCategory($keyword$)
e := expandConcepts($keyword$)
$exConcepts$:= {}
for each e_k in e
    Φ_{e_k} := getCategory(e_k)
    $\Theta_{e_k} = M_u(\Phi_o, T_{e_k})$
    $C'(e_k) = C(e_k) \cdot \Theta_{e_k}$
    $\gamma'(e_k) = \frac{C'(e_k)}{max(C'(e_k))}$
    $exConcepts$.add(e_k)
for each e_k in $exConcepts$
    $results$:= performSearch(e_k)
    for each r_i in $results$
        inst_score(r_i) := $G \cdot \sigma(r_i) + (1 - G) \cdot \gamma'(e_k)$
        addtoInstScores($inst\_score(r_i)$)
$scores[x] := \sum_{i=1}^{n} inst\_score(x_i)$
for each x in $scores$
    $relTags$:= getRelatedTags(x)
    for each r_k in $relTags$
        Θ_{r_k} := getCategorySimilarity(Φ_o, r_k)
    $scores'[x] := \frac{score[x] + \sum_{k=1}^{n} \Theta_{r_k}}{n+2}$
Sort by $scores'$ descending

Fig. 3. Algorithm

4. How much do you know about Web2.0 and Tags based web systems?
5. How easy to use, do you think, is the interface of the prototype?
6. Do you understand the concept of relations between concepts?
7. Do you find the concept of generality given in the prototype easy to understand?
8. Are you comfortable with the search system saving your search and/or access history?
9. Do you understand the problem of searching for content tagged with synonymous and/or polysemous words?
10. Have you ever experienced the above mentioned problems while searching for content on the web?
11. Do you understand the concept of common sense, specifically relating different concepts together?
12. Do you understand the technique used in this search system?
13. How would you rate the relevance of the search results to your query?
14. How would you rate the relevance of the search results to your intended target content?
15. Do you think the search results accurately depict your preference in ambiguous words?

16. Were there any irrelevant items in the returned results?
17. How would you rate the overall usefulness of the search results?

7 Results and Analysis

The results to the questionnaire are summarized in Table 2. Here, we briefly analyze the pattern in the results.

Persons	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	Answer Description
Questions									
4	3	1	2	1	1	3	1	2	1–4: Little knowledge – detailed knowledge
5	1	3	2	2	3	1	2	1	1–4: Easy – difficult
6	2	2	2	1	2	3	2	2	1–3: No understanding – complete understanding
7	2	3	3	2	3	2	3	1	1–4: Easy – difficult
8	1	1	1	2	2	2	3	1	1–3: Comfortable – not comfortable
9	2	2	1	2	3	1	3	2	1–3: Complete understanding – no understanding
10	2	3	2	3	3	1	3	3	1–3: Have experienced problems – have not
11	3	2	2	3	2	1	2	2	1–3: Clear – confusing
12	3	3	2	3	3	2	3	2	1–3: Understand – don't understand
13	2	3	2	4	2	2	3	2	1–4: Relevant – not relevant
14	2	3	3	3	2	2	3	2	1–4: Relevant – not relevant
15	3	2	2	3	3	2	3	3	1–3: Personalized – not personalized
16	2	3	2	3	2	2	3	3	1–3: No irrelevant results – many irrelevant results
17	2	3	2	3	2	1	4	2	1–4: Useful – not useful

Table 2. Summary of Results of the Effectiveness Study

Some of the important points to note in these results are the following:

- Answers to Question 7 – "Do you find the concept of *generality* given in the prototype easy to understand?" – suggest that users of the prototype found the concept of generality difficult to grasp. It seems therefore that this variable should be automatically adjusted by any system implementing our model instead of leaving it up to the users to pick its level. We do not think it would be appropriate to embed the value of generality in the model itself because it depends on the context of search and should be left customizable to the individual implementation.
- Several users found the graphical user interface of the prototype a little difficult to understand. While it was not our primary goal to make the prototype easy-to-use, an easier front-end might have shown better results in the effectiveness study. However, this finding does not affect the actual model.
- Many participants, in response to Question 12 – "Do you understand the technique used in this search system?" – answered that they did not understand the technique used in our prototype. In social networks, it is of

immense importance that the users understand the underlying reasoning mechanisms as much as possible. It helps them use the network more effectively. Any service implementing our model needs to put some efforts in educating the users about the working of intelligent search to enable them to utilize it more effectively.
- The issue of noise, according to responses to Question 16 – "Were there any irrelevant items in the returned results?" – was not effectively resolved by our prototype. We believe that the reason for this is that the participants of the survey did not have a detailed *user profile* in our prototype's database. Personalization depends heavily on this profile but it takes a little while to create an effective corpus for each individual user. We believe that with use, the effectiveness of the personalization module would increase. However, a proof of this cannot be obtained without an extensive study conducted over a long period of time on a larger number of constant users.
- It is evident from the answers to Question 8 – "Are you comfortable with the search system saving your search and/or access history?" – that privacy is not an issue in users of our geographical proximity. There seems to be a need to educate the users about privacy being an important issue which should be taken more seriously. However, it is an issue outside the scope of this research and is not our primary concern.

8 Future Work

Search results are, by nature, difficult to analyze and require users' subjective analysis. While the initial tests with the proposed technique of using personalized web search with common sense and folksonomy based search systems has shown positive results, a more detailed usability study is necessary to study the effectiveness of the technique for different users. Future work along this path aims to conduct detailed experimental studies using this new technique using real-world folksonomy based applications such as flickr [12] and Wordpress [13] etc. A comparison with other search techniques is also necessary to determine the full effectiveness of the proposed technique.

This technique still utilizes only three sources of information: tags, user profile and search engine's score. While these are the primary source of content's meta information in a folksonomy based service, other ranking variables, such as links to related content, are still not utilized. This technique may benefit from a more thorough study on how content clustering and relevance feedback techniques may be incorporated in this approach for better ranking of search results.

9 Conclusions

The information overload caused by the coming of user-created data on Web2.0 can only be addressed by utilizing all available resources for search and organization. User created organization of data has produced acceptable levels of results but still has problems because of variances in users creating this organization. A

possible solution to this problem is the application of machine common sense to the problem of search. In this research work we have outlined a framework for using the Open Mind Common Sense project to address the issue. This is done through the application of ConceptNet, a freely available tool kit for machine common sense on folksonomy.

The model, proposed in this research work, uses common sense and folksonomy and offers a different approach towards addressing the issue of search in social networks. However, it also leads to some noise in search results due to polysemy. To overcome this issue of noise, we enhanced the basic technique using a search results personalization technique. A detailed description of a modified approach for utilizing a personalized web search technique for returning more relevant search results in a CS&F based search system is described.

An effectiveness study was developed for measuring the success of the proposed approach. Different users, from different technical and non-technical backgrounds were asked to evaluate the prototype and give their opinions through a questionnaire. The results were collected and analyzed to measure the effectiveness of the prototype. The results have shown that while the prototype was able to demonstrate better recall, it has been prone to some noise in the results. This might be due to the reason that the participants of the study did not have an extensive search and access history in the system and the system was thus unable to perform personalization as effectively as it could have.

References

1. Golder, S., Huberman, B.: The Structure of Collaborative Tagging Systems. Arxiv preprint cs.DL/0508082 (2005)
2. Lieberman, H., Liu, H.: Adaptive Linking between Text and Photos Using Common Sense Reasoning. Conference on Adaptive Hypermedia and Adaptive Web Systems (2002)
3. Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., Minutoli, S.: SemKey: A Semantic Collaborative Tagging System. Proceedings of 16th International World Wide Web Conference, WWW2007 (2007)
4. Liu, F., Yu, C., Meng, W.: Personalized Web Search by Mapping User Queries to Categories. Proceedings of the eleventh international conference on Information and knowledge management (2002) 558–565
5. Nauman, M., Hussain, F.: Common Sense and Folksonomy: Engineering an Intelligent Search System. In: Proceedings of ICIET'07: International Conference on Information and Emerging Technologies, IEEE (2007)
6. Nauman, M., Khan, S.: Using Personalized Web Search for Enhancing Common Sense and Folksonomy Based Intelligent Search Systems. In: Proceedings of WI'07: IEEE/WIC/ACM International Conference on Web Intelligence. (November 2007)
7. Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., Zhu, W.: Open Mind Common Sense: Knowledge acquisition from the general public. Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems (2002)
8. Liu, H., Singh, P.: ConceptNet: A Practical Commonsense Reasoning Tool-Kit. BT Technology Journal **22**(4) (2004)

9. Singh, P.: The public acquisition of commonsense knowledge. Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access (2002)
10. OMCS: The Open Mind Common Sence Project. Accessed at: http://openmind. media. mit. edu/
11. Singh, P., Williams, W.: LifeNet: a propositional model of ordinary human activity. Proceedings of the Workshop on Distributed and Collaborative Knowledge Capture (DC-KCAP) at K-CAP (2003)
12. Flickr: About flickr. http://www. flickr. com/ about/ (Retrieved on February 24, 2007)
13. WordPress: Wordpress.com. Accessed at: http://www. wordpress. com/ (Retrieved on November 13, 2007)