# *Resolving Ambiguity Using Word Embeddings for Personalized Information Retrieval in Folksonomy Systems*

Ghazale Etemadikhou
Department of Computer Engineering, University of Science & Culture Tehran, Iran
computer.software1390@gmail.com

Fatemeh Azimzadeh*, Abdolsamad Keramatfar
Scientific Information Database (SID), ACECR, Tehran, Iran
Azimzadeh@acecr.ac.ir, samad@sid.ir

Received: 2020/10/28 Revised: 2021/03/31 Accepted: 2021/04/07

***Abstract*— The diversity and high volume of available information on the web make data retrieval a serious challenge in this environment. On the other hand, obtaining user satisfaction is difficult, which is one of the main challenges of data retrieval systems. Depending on their information about interests and needs for the same keyword, different people expect different responses from Information Retrieval (IR) systems. Achieving this goal requires an effective method to retrieve information. Personalized Information Retrieval (PIR) is an effective method to achieve this goal which is considered by researchers today. Folksonomy is the process that allows users to tag in a specific domain of information in a social environment (tags are accessible to other users). Folksonomy systems are made collaborative tagging systems. Due to the large volume and variety of tags produced, resolving ambiguity is a severe challenge in these systems. In recent years, word embedding methods have been considered by researchers as a successful method to fix the ambiguity of texts.**

**This study proposes a model which, in addition to using word embedding methods to remove tag ambiguity, provides search results in a personalized approach by fixing ambiguity and sentiment analysis combination tailored to users' interests. In this research, different models of word embeddings were applied. The experiments' results show that after applying the fixing ambiguity, the mean accuracy criterion improved by 1.93% and the mean** MRR (**Mean Reciprocal Rank) by 0.38%.**

***Keywords—Personalized Information Retrieval; Folksonomy; Fixing Ambiguity; Word Embedding; Sentiment Analysis.***

## 1. INTRODUCTION

One of the important goals of IR systems are to gain users' satisfaction. PIR was introduced with the aim of identifying the characteristics and interests of users so that among the many answers that can be provided per topic, the closest answer to the user's needs is provided to her/him.

One of the social networks which use the personalized information retrieval is the folksonomy system [1]. Users in this system have personal profiles and produce various tags for existing content. Therefore, it is difficult to control tags in terms of error, inconsistency, and ambiguity so, if errors and inconsistencies are not fixed in this system, data retrieval will encounter errors [2].

Fixing ambiguity is a very important issue in Natural Language Processing (NLP). Ambiguity is a barrier to machine language understanding. The large volume of tags in folksonomy systems and the lack of restrictions and control over tags are a factor in creating errors and ambiguity in the writing of tags. This problem reduces the efficiency of results retrieved for users. In order to solve this problem, previous researches have used the method of query expansion and embedding words. Tags in the folksonomy system contain users' thoughts and feelings, and by sentiment analyzing or mining opinion on tags, it is possible to identify users' interests. This can effectively help generate personalized responses to searches.

Some studies, such as Zhou et al. [2], have addressed the challenge of resolving the ambiguity of tags in the folksonomy system or, like Xie et al. [1], have focused on the challenge of sentiment analysis in the folksonomy system.

This study aims to improve the retrieval efficiency of PIR in the folksonomy system by fixing ambiguity with different word embedding models, on the other hand incorporating sentiment analysis in the suggested model. Word embedding models were seriously considered by the research presented by the Word2vec model [3, 4] and with a significant effect on increasing accuracy, it was an introduction to the development of other related algorithms.

This article is written in six sections. The second part reviews the research on fixing ambiguity and sentiment analysis, especially in folksonomy tags. The third part will describe the research method. In the fourth section, the experiments and the results related to the proposed approach will be analyzed. In the fifth section, conclusions are made and finally, in the sixth section, future works are presented.

This problem reduces the efficiency of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a journal. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

## 2. BACKGROUND

### *2-1 Fixing Ambiguity in Folksonomy Systems*

A review of studies conducted in the field of fixing ambiguity in folksonomy systems shows that one of the challenges are user profile enrichment. In many cases, user profile information is not enough to search because the users may not be very active in their profile. Therefore, external sources that are close to the user's interests have been suggested as a solution to the user's profile [2]. A serious drawback in this method is the change of the main identity of the users. In another approach [5], tag enrichment has been accomplished using two methods of Latent Dirichlet Allocation (LDA) and Wikipedia Category links. In fact, LDA examines the one-to-one relationships between tags. Wikipedia is used as an external resource that contains a large number of resources related to a specific topic.

The word2vec model was introduced in 2013 and took the issue of word embedding into language processing seriously. Using the word2vec method, semantic relationships between tags are obtained. According to this method, tags can be enriched, and ambiguity will be fixed [3, 4].

In the other approach, combining word embedding model with the WordNet semantic body, the words are examined semantically separately once with the Word2vec method and once with WordNet, the existing ambiguities are fixed then finally, based on the results obtained from both sections, ranking of the words is created [6].

Fernandez et al. [7] have proposed a resolving ambiguity method, in which all words are taken into the vector space then a similarity matrix is constructed. The obtained candidate words are taken to a two-part graph for enrichment, and finally, based on the points given, the word relationships are determined.

Fixing ambiguity and ranking resources is performed in four steps in the method proposed by Wang et al. [8]. These steps include; 1. Select the appropriate dataset. 2. Extract relationships, words in retrieved documents, and words with WordNet. 3. Similarity evaluations, words that have been ambiguity fixed in the previous step are vectored and the similarity between them is measured. 4. Semantic extraction, words and sentences are pre-processed and converted into matrices, the similarities between them are checked and finally, the ultimate ranking is performed.

### *2-2 Sentiment Analysis in The Folksonomy System*

Sentiment analysis is used as a tool for NLP processing. In other words, discovering and understanding the personal opinions of people, interests, and user feedback is called opinion mining, which is called sentiment analysis [9], too.

Sentiment analysis has five important sections: the extraction of entities, the classification of interests and sentiments, the popularity of entities, the management and maintenance of classes, and the timing of classifications [10].

According to research [9], opinion mining is accomplished based on five steps; these steps are 1. Making words vectors, which is accomplished based on the word embedding method. 2. Sentences are generated using the relationships between words. 3. At this stage, the input will be sentences and semantic relations between them then the output are entities related to the words and sentences based on opinion mining. 4. Each of the entities is taken to the vector space and ranked. 5. Each entity with the highest score represents beliefs about a subject.

Other researches such as [10, 11] have summarized different tools in the field of sentiment analysis and the latest approaches in the field of sentiment analysis, including sentiment polarity, TF-IDF model, and word embedding. Sentiment polarity expresses the sentimental aspects of ideas. In the texts, sentimental results are specified for each part of the sentence. Sentiment polarity has positive, negative, and neutral weights. The TF-IDF model is a statistical model that shows how important a word is to a document in a set of documents.

In a study conducted by Shi et al. [1], user profiles and resource profiles are created, and then a matrix of user-tag and resource-tag relationships is constructed. Using the SenticNet library, relationship matrices are sentimentally analyzed.

### *2-3 Sentiment Analysis and Word Embedding Combination Methods*

The purpose of this study was to fix ambiguity and sentiment analysis to retrieve personalized information in folksonomy systems. Since the creation of tags in this system is performed collaboratively by users, so its production and publication are not subject to specific rules, which make lots of variety in tags; therefore, fixing ambiguity is very important. Since the research presented in this paper demonstrates the effectiveness of word embedding models to fix the ambiguity, this research intends to use the word embedding model to fix ambiguity.

PIR also means taking into account the tastes of individuals in retrieving information. In other words, among the available answers, the closest answer to the individual's taste receives more scores for the query phrase. Opinion mining is one of the methods that allow identifying users' tastes; therefore, to personalize the results, this study identifies users' tastes using sentiment analysis methods.

Obviously, a combination of fixing ambiguity and sentiment analysis methods can be used to retrieve information. Since tags are very important in folksonomy systems and express the thoughts and feelings of users and ambiguities that occur in these tags, we intended to use a combination of two methods of fixing ambiguity and sentiment analysis in the folksonomy system to improve user satisfaction.

## 3. RESEARCH METHOD

The folksonomy system is one of the best systems that can be used for PIR. In this system, tags are created by users for the existing content in the system, which themselves serve as a powerful tool in the data retrieval model, but since there are no specific instructions for its production, it is necessary to remove their ambiguity for better efficiency. Personalizing information retrieval increases user satisfaction, and as stated in the previous section, resolving

ambiguity and sentiment analysis are effective tools to achieve this goal. Fig.1 shows the steps of performing different activities in this research.

*3-1. DataSet*

The dataset used in this study is Movielens, which has been used in other studies, for instance [1, 2, 12]. In this study, Movielens[1] dataset was used. This dataset is a free dataset and includes movies and their metadata include; 71,567 users, 10,681 videos, 95,580 tags, and 100,000,56 scores. In this system, user profiles and resource profiles are created based on tags.

*3-2. Baseline System*

First, we need to implement the Baseline. The basic model is actually the folksonomy system, which is created from the users, resources and tags, also the triple relationships between them. This triple relationship helps to create user profiles and resource profiles [1].

Equations (1) and (2) show the relationships in the folksonomy system, which U is the set of users, R is the set of resources, T is the set of tags and K is the set of relationships between these three sets,

$$\theta = (U, R, T, K) \quad (1)$$

$$K \subseteq U \times R \times T \quad (2)$$

According to the relation mentioned above, θ is divided into three parts to rank the sources.

Equations (3), (4), and (5) show the resources scoring which Q is set of queries; S1 is the score between tags and resources that actually indicates the content relationship between resources. S2 is the score between resources and users, which indicates the same interest of users in resources, and S is the result of combining the two scores S1 and S2, which is the final score (S) for ranking of resources based on user's interest and query.

$$\theta_1 = R \times Q \rightarrow S_1 \quad (3)$$

$$\theta_2 = R \times U \rightarrow S_2 \quad (4)$$

$$\theta_3 = S_1 \times S_2 \rightarrow S \quad (5)$$

To apply the model for the folksonomy system, we first need to create a user profile. Assume that $\{t_1, \dots, t_n\}$ is a set of tags used by a Ui user and $\{P_1^i, \dots, P_n^i\}$ is the number of times a user has used a tag.

Equation (6) shows the user profile in the folksonomy. Pn is obtained from different methods, including NTF (Normalized Tag Frequency) or other term frequency-based models [13], [14], [15].

$$U_i = (t_1 \rightarrow P_1^i, \dots, t_n \rightarrow P_n^i) \quad (6)$$

In this research, we have used the

[1]http://movielense.org/datasets/movielense/

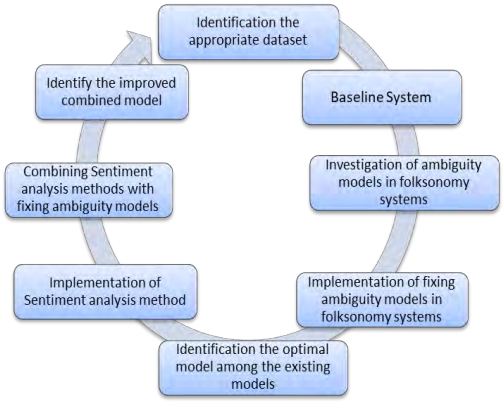


Fig. 1: Implementation Steps

NTF method, the following equations are used to create user and resource profiles, where in (7), $N_{i,x}$ is the number by which the i[th] user assigns the x tag to resources, and $N_i$ represents the number of resources tagged by users. Moreover, in (8), $M_{c,x}$ indicates the number of users who use the x tag in resource c, and $M_c$ is the total number of users who have tagged resource c.

$$U_{i,x} = \frac{N_{i,x}}{N_i} \quad (7)$$

$$W_{c,x} = \frac{M_{c,x}}{M_c} \quad (8)$$

The same method is used to create resource profiles as was used to create user profiles. Assume that $\{t_1, \dots, t_n\}$ are a set of tags given to the source $R_a$ and $\{N_1^a, \dots, N_m^a\}$ the number of times these tags are repeated on this source. Equation (9) shows the resource profile.

$$R_a = (t_1 \rightarrow N_1^a, \dots, t_n \rightarrow N_m^a) \quad (9)$$

Once the user and resource profiles are created, we convert each of the two profiles into a matrix. In the $W_u$ users' matrix, each row is a user, and each column is a tag, and in the $W_r$ resources matrix, each row is a source and each column is a tag.

$$W_u = \begin{bmatrix} P_1^1 & P_2^1 & \cdots & P_n^1 \\ P_1^2 & P_2^2 & \cdots & P_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ P_1^i & P_2^i & \cdots & P_n^i \end{bmatrix}$$

$$W_r = \begin{bmatrix} N_1^1 & N_2^1 & \cdots & N_m^1 \\ N_1^2 & N_2^2 & \cdots & N_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ N_1^a & N_2^a & \cdots & N_m^a \end{bmatrix}$$

In order to rank resources, it is necessary to determine the similarity between resources with users and the similarity between resources with tags, which (10) and (11) respectively show similarity of users with resources and similarity of the query with resources, so (12) show

calculates final ranking score of similarity between query and resources, based on user interest.

$$\mathrm{Sim_c(U_i, R_a)} = \frac{\mathrm{U_i . R_a}}{||\mathrm{U_i}||\mathrm{R_a}||} \quad (10)$$

$$\mathrm{Sim}_c(q, R_a) = \frac{q . R_a}{||q| . |R_a||} \quad (11)$$

$$(U_i, q, R_a) = e^{Sim_c(U_i, R_a) + Sim_c(q, R_a)} \quad (12)$$

Each resource has a score and rank; higher-ranked resources are retrieved as a search result.

In this stage, the original folksonomy system (Baseline System) has been created, which considers the similarity between resources and query, also user personality interest. Now we fix the ambiguity on this system to improve the performance of data retrieval.

### *3-3. Fixing Ambiguity in The Folksonomy System by Word Embedding Methods*

This research used the word embedding method to fix ambiguity. In the word embedding method, each word is mapped to a real-valued vector in a vector space, such that words with close semantic meaning have vectors close to each other. Word embedding helps to fix ambiguity. Word embedding is a technique for learning linguistic features and a set of language modeling used to synthesize words. This research used Word2vec [3], [4], fastText and GloVe models.

Word2vec is a model for embedding words. This algorithm takes a piece of text as input and maps it to a vector space. Word vectors are located in this space, and words that have a semantic meaning close to each other have close vectors in this space.

The code used in this algorithm is simple but has a strong architecture. Word2vec works in any volume of the dataset and does not use much memory to execute code [16].

fastText: In machine learning, each piece of data is tagged and classified accordingly. FastText is another embedding model, which uses supervised and unsupervised algorithms to model words.

GloVe: A global vector model that combined the count-based matrix factorization model and the content-based skip-gram model.

After implementing the baseline system for the folksonomy system, three Word2vec, fastText, and GloVe models are run on it.

### *3-4. Analysis Of Sentiment in The Folksonomy System*

SenticNet is about the level of sentiment analysis. In general, it is the recognition, discovery, and orientation of sentiments using useful and conceptual information related to words and terms and the number of repetitions of them in a text. SenticNet is used for sentiment analysis, which is a structure for calculating and analyzing the sentiments of words. Words are given sentimental values and based on these values, it is determined which words are close to each other in terms of sentimental meaning [17].

In this research, we intended to increase the information retrieval performance in the folksonomy system by combining both fixing ambiguity and Sentiment analysis models. The study focused on resolving ambiguity on the model, which considered PIR.

## 4. Tests and Results

Fixing ambiguity in the folksonomy system improved criteria such as P@N and MRR in evaluation. The measure of Precision is equal to the fraction of retrieved documents that are related to the requested information. P@N is the number of correct retrieved answers in the N documents. The MRR is an evaluation criterion that indicates the probability of the correct answer based on the ranking of the final results.

In order to evaluate the proposed models, we need to compare the changes made in the results retrieved by the models compared to the other models. For this purpose, the results of the fixing ambiguity and sentiment analysis section are examined separately.

Fixing ambiguity models were used separately with the same test conditions. Table 1 shows the test results in the system (Baseline System) and word embedding models. The results in this table show that; "Baseline with Pre-trained word2vec", "Baseline with fastText" and "Baseline with GloVe" were improving all criteria, which "Baseline with fastText" tackled the best.

Therefore, according to the resolving ambiguity models in folksonomy, and the performance improvement of the models in ambiguity, the models are shown in Table 1, respectively. Among the various models for debugging, Model 8 has better performance in debugging the folksonomy system.

In order to improve data retrieval in the folksonomy system, a combination of two methods of ambiguity and sentiment analysis in tags was performed. Sentiment analysis is performed on eight models of fixing ambiguity that is shown in Table 2.

The combination of "Baseline and fastText" achieved the best results in the fixing ambiguity phase as well as after the implementation of sentiment analysis but incorporating sentiment analysis is improved "Baseline with Trained word2vec 10", that is number 7 in the table above.

TABLE 1. Results of fixing ambiguity algorithms

| NO | Models | P@5 | P@10 | P@15 | P@20 | MRR |
|---|---|---|---|---|---|---|
| 1 | Baseline system | 15.67 | 22.31 | 26.15 | 28.78 | 10.89 |
| 2 | Pre-trained Word2vec | 7.23 | 11.59 | 14.89 | 17.40 | 5.31 |
| 3 | Trained word2vec 10M | 8.63 | 13.85 | 17.51 | 20.33 | 6.03 |
| 4 | fastText | 8.19 | 13.11 | 16.52 | 19.50 | 5.92 |
| 5 | Glove | 7.66 | 12.16 | 15.55 | 18.36 | 5.47 |
| 6 | Baseline with Pre-trained word2vec | 16.06 | 23.68 | 28.34 | 31.35 | 11.14 |
| 7 | Baseline with Trained word2vec 10M | 12.96 | 19.18 | 23.12 | 25.96 | 8.96 |
| 8 | Baseline with fastText | 16.25 | 24.07 | 28.62 | 31.68 | 11.27 |
| 9 | Baseline with GloVe | 16.16 | 23.82 | 28.23 | 31.33 | 11.20 |

TABLE 2. RESULTS OF FIXING AMBIGUITY AND SENTIMENT ANALYSIS COMBINATION

| NO | Models | P@5 | P@10 | P@15 | P@20 | MRR |
|---|---|---|---|---|---|---|
| 1 | Baseline system | 15.76 | 22.47 | 26.28 | 28.91 | 10.87 |
| 2 | Pre-trained Word2vec | 6.82 | 11.20 | 14.17 | 17 | 5.10 |
| 3 | Trained word2vec 10M | 8.36 | 13.49 | 16.83 | 19.50 | 5.91 |
| 4 | fastText | 8.21 | 13.11 | 16.54 | 19.51 | 5.92 |
| 5 | Glove | 7.47 | 12 | 15.53 | 18.35 | 5.42 |
| 6 | Baseline with Pre-trained word2vec | 15.86 | 23.48 | 27.93 | 30.9 | 10.73 |
| 7 | Baseline with Trained word2vec 10M | 16 | 22.89 | 26.98 | 29.87 | 10.89 |
| 8 | Baseline with fastText | 16.65 | 24.02 | 28.52 | 31.64 | 11.62 |
| 9 | Baseline with GloVe | 16.43 | 23.81 | 28.15 | 31.22 | 11.52 |

## 5. CONCLUSION AND FUTERE WORKS

We achieved the goal of this study as we improved the performance of the baseline system by combining word embedding models with baselines.

In this study, we tried reducing ambiguity using word embedding models; the "Baseline with fastText" model has the best performance in Precision and MRR criteria. On the other hand incorporating sentiment analysis was improving "Baseline with Trained word2vec 10".

The paper more focused was on PIR and applying different word embedding models for resolving ambiguity in folksonomy systems after that try to incorporated sentiment analysis on the model. For future study, may be consider different model of sentiment analysis to improve the evaluation criteria.

Although the proposed method has a good performance, other suggestion for future works are; 1. Combining the WordNet method with word embedding models helps to fix the ambiguity of tags precisely in the folksonomy system. 2. Enrich user profile: The user has no activity in the system, so no information is available about the user's interests to be retrieved based on the user's interests.

## REFERENCES


[1] H. Xie et al., "Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy," *Information Processing & Management*, vol. 52, no. 1, pp. 61-72, 2016.

[2] D. Zhou, X. Wu, W. Zhao, S. Lawless and J. Liu, "Query expansion with enriched user profiles for personalized search utilizing folksonomy data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1536-1548, 2017.

[3] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neuralinformation processing systems*, 2013, pp. 3111-3119.

[5] I. Saleh and N. El-Tazi, "Finding semantic relationships in folksonomies," In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, 2018, pp. 174-181.

[6] N. Pittaras, G. Giannakopoulos, G. Papadakis, and V. Karkaletsis, "Text classification with semantically enriched word embeddings," *Natural Language Engineering*, pp. 1-35, 2020.

[7] F. C. Fernández-Reyes, J. Hermosillo-Valadez, and M. Montes-y-Gómez, "A prospect-guided global queryexpansion strategy using word embeddings," *Information Processing & Management*, vol. 54, no. 1, pp. 1-13, 2018.

[8] Y. Wang, M. Wang, and H. Fujita, "Word sense disambiguation: A comprehensive knowledge exploitation framework," *Knowledge-Based Systems*, vol. 190, p. 105030, 2020.

[9] A. D. Vo, Q. P. Nguyen, and C.-Y. Ock, "Semantic and syntactic analysis in learning representation based on a Sentiment analysis model," *Applied Intelligence*, vol. 50, no. 3, pp. 663-680, 2020.

[10] O. Habimana, Y. Li, R. Li, X. Gu, and G. Yu, "Sentiment analysis using deep learning approaches: an overview," *Science China Information Sciences*, vol. 63, no. 1, pp. 1-36, 2020.

[11] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, 2020.

[12] T. Luo, S. Chen, G. Xu, and J. Zhou, *Trust-based collective view prediction*, Springer, 2013.

[13] Y. Cai and Q. Li, "Personalized search by tag-based user profile and resource profile incollaborative tagging systems," In *Proceedings of the 19th ACM international conference on information and knowledge management*, 2010, pp. 969-978.

[14] M. G. Noll and C. Meinel, "Web search personalization via social bookmarking and tagging," In *The semantic web*, Springer, 2007, pp. 367-380.

[15] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval*, 2008, pp. 155-162.

[16] X. Rong, "word2vec parameter learning explained," arXiv preprint arXiv:1411.2738, 2014.

[17] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: a common and common-sense knowledge base for cognition-driven Sentiment analysis," In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.


**Ghazale Etemadikhu** is a graduate student in master of software engineering at the University of Science and Culture, Tehran, Iran. Her research interest is information retrieval.

**Fatemeh Azimzadeh** received a Ph.D. degree in Information Technology from University Putra Malaysia in 2012. Currently, she is an assistant professor in ACECR, Tehran, Iran. She is also the director of SID (Scientific Information Database) in Iran. Her research interests include information retrieval and information quality.

**Abdalsamad Keramatfar** is a Ph.D. student in information technology at the University of Qom and a data scientist at SID. He currently works in natural language processing and machine learning and specifically on multi-thread modeling of context for social media sentiment analysis. His research interests are artificial intelligence and natural language processing.