Synonyms Detection in Folk Tag Set: A Novel Hybrid Solution

Nabila Amir✉*, Nabila Rehman, Reema Nawaz, Fouzia Jabeen*

Department of Computer Science,
Shaheed Benazir Bhutto Women University,
Peshawar, Pakistan

ABSTRACT

Collaborative tagging is one of the most important applications of web 2.0 that allow users to associate tags (a free-form text chosen by the users) with the resource, which is metadata for that resource. These tags are used later on for search and retrieval of these resources. One of the issues in a folk tag set is ambiguity, as ambiguity causes incorrect resource(s) retrieval. To bring precision in search, we need to remove this ambiguity. One of the reasons of ambiguity is presence of synonyms in a tag set. In this work, we have proposed a novel solution for synonyms detection. The proposed solution provides a concise tagset that will be associated with the resource. The methodology of our approach can be defined in four major steps. First, we have removed misspelled tags. In the second step, we have detected synonyms using WordNet and Microsoft Word dictionaries. In the third step, we have used Euclidian distance to find rest of the synonyms and finally, we obtained precise tag set without synonyms. Dictionaries provide coverage to tags which are Standard English language words and mathematical formula covers the tags which are from folk vocabulary and are not present in the dictionaries. We have tested our approach on image resources with which tag set composed of twenty tags is associated. We compared our results with five state-of-the art techniques including cosine, Jaccard, projection, mutual information, and dice. We can conclude that the results of our approach are more accurate in finding synonyms.

Keywords: *Tags, Folksonomy, Synonyms, Collaborative tagging, Search, Retrieval*

1. INTRODUCTION

Web 2.0 has introduced many social applications such as blogs, social networks, social bookmarking, photo, music and video sharing [1]. Web 2.0 is considered as second version of web. The important feature of this version of web is its flexible relation with the users.

Collaborative tagging is one of the most important applications of web 2.0. Collaborative tagging is also called group tagging as many users assign tags to a resource, and it is an activity that enable users (group of folks) to add tags to resource(s). These tags act as

✉nabilaameer1@gmail.com

metadata for that resource. Tags are freely chosen keywords assigned by the users. In response to collaborative tagging activity, folksonomy (stand for "folk taxonomy") is created. In other words, folksonomy is vocabulary of users. In folksonomy based systems, classification of resources is done according to users' vocabulary (Tags). Del.icio.us (Del.icio.us is a tagging system for URLs that is incorporated with Firefox browser and allows bookmarking. Bookmark lets and enable users to collect and recover their bookmarks on the del.icio.us website and display their bookmarked URL through tagging with suitable metadata) and Flickr (an online photo storage system that allows users to classify their photographs by tagsets) are two most popular examples of folksonomy systems.

In contrast to folksonomy, taxonomy is such science or technique in which the classification of the things and concepts is based on some principle. Folksonomy has some prominent features which make it popular nowadays. In folksonomy based systems, users can add large number of tags to a resource to describe the concept of the resource. The folk tag set built as a result of this activity improves the understanding of the resource as users utilize their own vocabulary. Tags are easy for everyone (belongs to different age, language, and belongs to different cultures) to remember and to utilize later for search and retrieval. Users that assign the same tag(s) to a resource mean that both have same interest and will be creating communities. Users can assign any number and combination of tags to express a concept, therefore folksonomies are multi-dimensional. Tags reproduce the user's conceptual model without cultural, social, or political bias. Folksonomy is flexible because users can add and remove tags [2].

When multiple users assign tags to a resource in which they show common interest toward that resource, synonyms may appear in a tagset. The presence of synonyms in a tagset causes ambiguity (a word or expression that can be understood in two or more possible ways). As a result, incorrect semantic interpretation of a resource may occur. In addition, it will give incorrect search and retrieval results. To solve this problem, we proposed a hybrid approach for synonyms detection in order to get a precise tagset where the term hybrid is used in the sense that we have utilized dictionaries and mathematical formulas in combination.

Our research work provides benefit to users by suggesting a precise folk tagset without synonyms. It will reduce confusion of users in understanding a resource. In addition, it will improve accuracy in search and retrieval of that resource.

2. RELATED WORK

To keep discussion on state-of-the-art approaches in the literature for synonym detection precisely, we have classified them into dictionary based and mathematical formulas based approaches.

Detection of synonyms in [3] based their results on overlap in translation of semantically similar words in many bilingual dictionaries. Bilingual and monolingual (an English dictionary and corpus) dictionaries for the extraction of synonyms are utilized in [4] while only monolingual corpora is exploiting in [5] and bilingual corpora in [6] for finding synonyms. The accuracy of monolingual is high but the coverage is low. However, it extracts a limited number of synonyms. The synonyms detection task is improved in [7] using the methods of [5][6]. They concluded that the results are better than use of individual methods.

WordNet (semantic lexicon for English language) is utilized in [8] to group words of English into group of synonyms. A test called WordNet based similarity test (WBST) is proposed in [9] . This test gives a large number of questions related to Test of English as a Foreign Language (TOEFL). WBST has 23,570 more questions as compared to TOEFL. WBST give much better result (72.2%) estimates than TOEFL. The Extended WordNet based similarity test (EWBST) is proposed in [10]. By using same and related words as distractors, EWBST enhances the WBST. EWBST is much complex than WBST but is more accurate in synonyms detection than WBST.

WordNet is represented as a graph G = (V, E) in [11]. In this graph, nodes of graph represents the WordNet concepts (synsets) and dictionary words; undirected edges denotes relationship between synsets and dictionary words are linked to the synsets related to them by directed edges. In the pair for each and every word, first authors [11] have calculated a personalized PageRank vector of graph G. They concentrated all probability mass in the target word. Concerning implementation of PageRank details, they chose a damping value of 0.85 and ending the task of calculation after 30 repetitions. These values are kept default values in their adopted simpler similarity method.

Tag to tag and resource to resource similarity is focused in [12]. For this purpose, authors have used various methods of aggregation such as distribution, projection, incremental, and collaborative filtering. They evaluated their results against cosine, overlap, Jaccard, and mutual information. Aggregation measures and distributional information give better result as compared to projection while the mutual information is observed as most precise measure for tag similarity. After matching the results of overlap, Jaccard and Dice, the authors concluded that these three approaches have no differences in meaning. Micro-aggregation is observed as the worst aggregation process, whereas, collaborative aggregation gives better result.

To identify similar tags, a technique called Adaptive Jenses-Shannon Divergence (AJSD) is proposed [13]. They evaluated their technique with cosine similarity and concluded that it gives better results. Latent semantic analysis (LSA) is utilized in [14] to detect synonyms in TOEFL test. Their work is 64.4% truly accepted however, work done by

[15] and [16] is much better in accuracy which is 72.5% and 81.25% respectively but the algorithm (LSA) used in [15] and [16] is simple and is applied on longer corpus of 4.7 million words.

Morpho-syntatic and semantic similarity is focused in [17]. The authors have utilized Levenshtein distance for morpho-syntatic similarity, tag signature and cosine similarity for semantic similarity. The two methods (morpho-syntatic and semantic similarity) in combination are utilized. A principle called Mutual reinforcement is highlighted in [18][11]. The principle states that "two tags are thought to be similar if they have been related to a similar resource, and two resources are thought to be similar if they have been named with similar tags". The proposed methods in [19] called the continuous bag-of-words model (CBoW), and skip-gram model (SG) which can find similarity not only between single words but also between pair of words for example between king, queen and man, women Tag Context Similarity (TCS) with the combination of Resource Context Similarity (RCS) are Used for the synonym identification task in [20]. The most prominent point is that this technique does not depend on any variation to which the tag might be subjected (e.g., study and studies). For a given tag set, TCS is the simplest method for finding similarity because for this external knowledge is not required.

A technique called Edit Distance (ED) which is also named as Levenshtein Distance is proposed in [21]. They used Edit Distance for finding words with spell variations like work and works, and also able to detect similar words having special character between them for example, case-study and case study.

To find similarity between tagsets, the authors of [22] utilized the normalized version of Tag Context Similarity (TCS). They follow a method which is implemented in two major steps: first is problem reduction, which simplify analyses and calculation. Secondly, synonym analysis, which meant for detecting and analyzing the synonyms of a given tagset. However, this approach is independent of specific type of variation in a tag.
Hybrid solutions are found producing more accurate results as compared to dictionaries or mathematical formulas based approaches.

3. METHODOLOGY

In this section, we have discussed the methodology of our proposed hybrid solution for synonyms detection. The steps of the opted methodology are illustrated in Fig. 1.

A. *Tagset*

In the first step, original tagset is generated. This tagset contains tags given by users to a resource.

B. *Spell Checking*

In second step, spells of all the tags present in the tagset generated at first step are checked. We have discarded tags with incorrect spells. The spell checker uses Microsoft word spell checker facility.

C. *WordNet and Microsoft Word Dictionaries*

In the third step, two dictionaries that are WordNet and Microsoft Word are used in combination for synonyms detection in folk tag set.

WordNet is an English lexicon that groups the words of English into set of synonyms called synset. As WordNet is an electronic database, therefore it can also be accessed through Web browser. WordNet is used heavily in folksonomy related research problems however, for synonym detection, it faces issue of very low coverage. To cover this limitation, we have exploited Microsoft Word dictionary as well.

Microsoft Word has a well-built ability for precise synonym detection. The additional feature of this dictionary is its dynamic nature. Users can add new words, which makes its coverage better especially for folksonomy applications.

D. *Euclidean distance*

In fourth step, Euclidean distance (E-distance) is exploited to find synonyms. In mathematics, Euclidean distance is used for finding distance between two points, where these points can be in different dimensional space.

We have used this measure to find the similarity between tags. E-distance (x,y) gives minimum value if there is highest similarity between tags x and y. The increase in value shows least similarity.

Cosine similarity is generally used as a metric for measuring distance when the magnitude of the vectors is significantly uneven. Usually, we use cosine similarity with large text because using this distance matrix on paragraphs of data is not recommended, that is why we use Euclidean distance for tag similarity instead of cosine similarity.

Finally, we get refined set in which there are no synonyms. This obtained tagset is more precise as compared to the original one. In the next section IV, we have presented our experimental results.

4. EXPERIMENTAL RESULTS

We have performed experiments on three tagsets, each having twenty tags given to images (resource) of mobile, computer and car. These tagsets are generated by collaborative tagging activity of folks.

In Table 1, we have tabulated the results. It shows the original tagset given by the users, the refined tagset generated by our proposed approach, synonyms detected by dictionaries and that by Euclidean distance.

In Table 2-8, we have presented results of finding similarly between any tag pair by Euclidian distance and with five state of the art approaches which include Cosine similarity, Jaccard, Dice, Projection and Mutual information. We can conclude that Euclidean distance is more precise in finding similarity between tags as compared to the rest of five approaches.

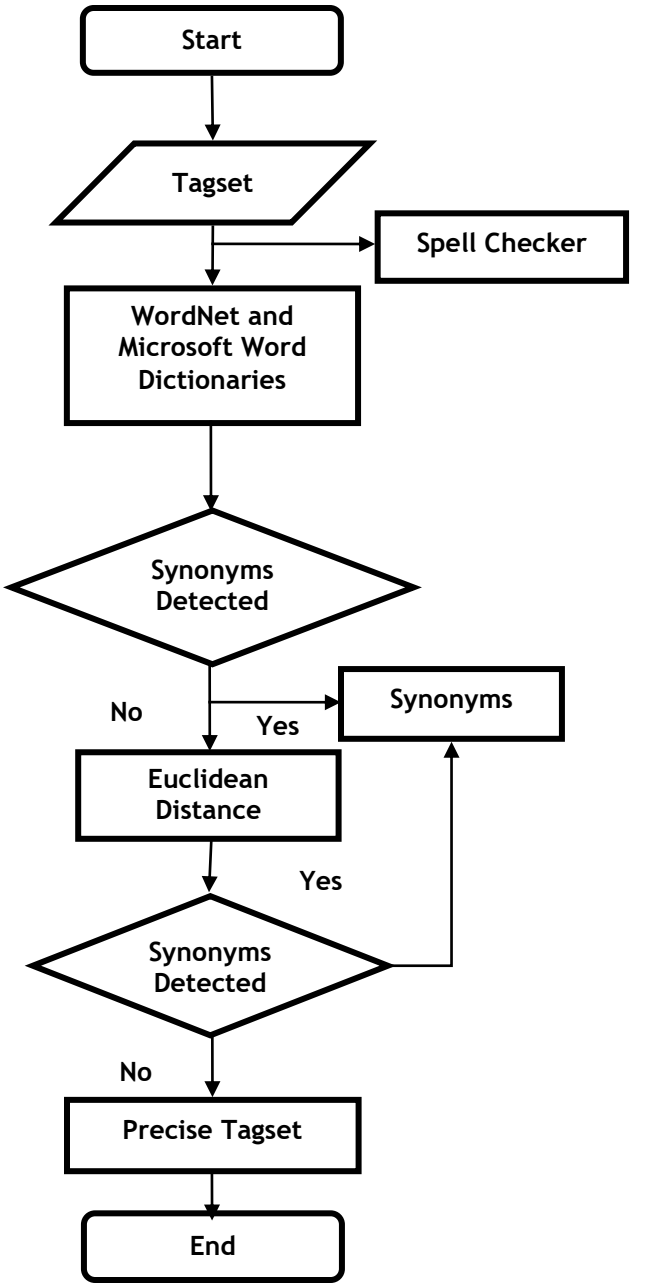

Fig. 1. Proposed Approach Framework

Table 1. Synonyms detection by dictionaries and Euclidean distance

	Original Tag Set	Refined Tag Set	Synonyms Detected by Dictionaries	Synonyms Detected by Euclidean Distance
1	movie, film, program, moving picture, snap, show, flick, video, picture show, big screen, silver screen, picture	Movie, flick, film, moving picture, video, big screen	Movie→film video→moving picture video→movie flick→picture	movie→film video→movie
2	Bag, carrier, purse, transporter, bucket, container, sack, shopping bag, carrier bag, paper bag, hand purse, wallet, pouch, pocket, small bag, money bag	Bag, carrier, purse, pouch, wallet, pocket, shopping bag, sack, bucket	Bag→purse purse→wallet carrier→transporter	bag→purse pocket→wallet
3	carriage, auto, automobile, compartment, wagon, coach ,flatcar, Pullman, bus, machine, car, cart, truck, train, van, lorry, motor, transport, carry	carriage, compartment, wagon, coach, flatcar, Pullman, bus, machine, car, truck, train, van, lorry, motor, transport	carriage→auto carriage →automobile automobile →cart	automobile →carry

Table 2. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approaches

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual Information
movie→film	1.001236e+002	4.442295e-001	2.212500e+004	7.011424e-001	9.379140e-001	63301
movie→program	1.121458e+002	3.999320e-001	1.843700e+004	7.206107e-001	5.697194e-001	17536
movie→moving ***picture***	2.166942e+002	1.142380e-001	2.234400e+004	7.811435e-001	9.934433e-001	48944
movie→snap	1.1001674e+002	3.139283e-001	4.684111e+004	9.056946e-001	8.173913e-001	26840
movie→show	1.111246e+002	6.529651e-001	44678	7.203946e-001	9.312503e-001	55104
movie→flick	2.0275335e+002	4.040164e-001	5.474600e+004	8.087966e-001	9.194895e-001	64746
movie→vedio	1.082557e+002	4.157136e-001	2.344600e+004	9.796861e-001	8.163615e-001	22609
movie→picture show	1.812275e+002	6.133366e-001	33726	7.133431e-001	8.171042e-001	43900
movie→big screen	2.489094e+002	9.425520e-001	9.995000e+004	8.194399e-001	7.853395e-001	45750
movie→sliver screen	2.4321579e+002	6.320376e-001	55006	5.643566e-001	9.081103e-001	22114
movie→picture	1.11224e+002	9.077586e-001	4708	7.098491e-001	9.180120e-001	67088
movie→photographs	1.255475e+002	6.555703e-002	5.559300e+004	6.094176e-001	9.950473e-001	56593

Table 3. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approaches

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual Information
film→program	1.432830e+002	1.234567e+001	5.939100e+004	1.104729e+001	5.218842e-001	47300
film→moving picture	1.000222e+002	2.661004e-001	73471	6.067942e-001	9.330126e-001	12391
film→ snap	1.223456e+002	3.388395e-001	40513	9.665512e-001	7.655893e-001	83471
film→ show	1.46779e+002	5.645745e-002	82000	5.103469e-001	9.983555e-001	60513
film→flick	2.530888e+002	7.32465e-001	8.00710 0e+004	1.345403e-001	8.616845e-001	92842
film→video	2.661525e+002	2.003328e-001	99682	9.220196e+000	7.646278e-001	0
film →picture show	2.322498e+002	5.566973e-001	54690	6.884528e-001	7.657001e-001	55382
film →big screen	1.244678e+002	6.451998e-001	86265	6.979274e-001	8.411702e-001	23394
film →sliver screen	3.163858e+002	9.456055e-002	73412	8.880565e-001	9.955325e-001	76265
film→picture	1.222245e+002	5.345009e-001	16114	9.313741e-001	8.605230e-001	33345
film →photographs	1.173073e+002	3.935028e-001	8.30010 0e+004	8.141189e-001	9.235716e-001	51526

Table 4. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approaches

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual Information
Program→moving picture	2.233644e+002	1.100256e+001	1.505900e+004	2.739294e-001	5.782552e-001	86570
program→snap	2.124730e+002	9.777204e-001	15401	1.062448e-001	6.858227e-001	56004
program→show	2.442785e+002	1.011853e+001	1.611200e+004	1.996677e-001	6.003679e-001	11660
program→flick	1.6278740+002	9.490321e-001	1.600600e+004	1.132509e-001	5.423797e-001	5218
Program→video	3.547251e+002	1.679466e-001	4312	2.346709e-001	4.884566e-001	19842
program→picture show	1.102526e+002	8.333369e-001	12234	1.491805e-001	5.846554e-001	7967
Program→big screen	3.44585e+002	1.115669e+001	1.796700e+004	2.101023e-001	4.161752e-001	76168
Program→silver screen	2.46238e+002	1.064827e+001	1.616500e+004	1.404278e-001	6.668201e-001	17384
program→picture	1.130259e+002	8.167762e-001	56980	1.722884e-001	4.628931e-001	17437
program→photograph	3.777456e+002	1.137238e+001	1.584700e+004	1.907065e-001	6.151036e-001	11004

Table 5. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approaches

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual Information
Moving picture→snap	1.457890e+002	7.039927e-001	1.637700e+004	7.018664e-001	7.776931e-001	29014
Moving picture →show	2.500135+002	3.825368e-001	35792	7.241053e-001	8.522639e-001	12589
Moving picture →flick	1.224377e+002	4.803032e-001	88550	1.147365e-001	9.277207e-001	89567
Moving picture →video	1.426374+002	7.022900e-001	36789	1.001302e+001	8.868549e-001	11034
Moving picture →picture show	1.006792e+002	7.007965e-001	2.167900e+004	7.246389e-001	7.633649e-001	78654
Moving picture →big screen	1.150186e+002	6.729313e-001	72134	8.173181e-001	7.643288e-001	64479
Moving picture →silver screen	1.245992e+002	4.307906e-001	1894400e+004	5.765250e-001	7.819980e-001	90134
Moving picture →picture	1.334467e+002	4.854102e-001	6.772200e+004	7.268968e-001	9.086359e-001	77231
Moving picture →photograph	1.538343e+002	5.337454e-001	4.976200e+004	7.001235e-001	8.844836e-001	67722

Table 6. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approaches

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual Information
snap→show	1.012568e+002	8.887232e-001	47012	3.333746e-001	9.147517e-001	12467
snap→flick	1.168930e+002	5.134288e-001	2.778000e+004	5.001512e-001	8.834617e-001	73410
snap →video	2.333478e+002	7.223150e-001	4.571000e+004	3.679006e-001	6.234107e-001	99012
snap→picture show	1.57924e+002	5.347207e-002	4.267100e+004	5.102884e-001	8.205707e-001	23761
snap → big screen	1.111865e+002	4.840093e-001	46125	3.553428e-001	3.986310e-001	10125
snap →silver screen	1.678910e+002	4.678232e-001	62190	5.324339e-001	3.289245e-001	48130
snap →picture	1.026596e+002	7.554701e-001	4.545900e+004	6.924988e-001	7.279493e-001	45459
snap →photograph	1.567193e+002	4.532748e-001	4.833900e+004	3.952046e-001	9.047198e-001	87901

Table 7. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approaches

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual Information
show→flick	1.499200e+002	5.669182e-001	4.26300e+004	3.491714e-001	5.866251e-001	47913
show →video	2.a+002	7.178703e-001	2.253700e+004	1.140297e+001	8.435600e-001	48524
show →picture show	1.831147e+002	7.158085e-001	43825	1.031193e+001	7.532083e-001	54263
show →big screen	1.873339e+002	5.731409e-001	80853	9.003530e-001	7.545630e-001	22537
show →silver screen	3.114482e+002	1.108631e-001	7.804400e+004	6.545691e-001	8.402019e-001	43825
show →picture	1.507813e++002	5.704121e-001	56549	8.326777e-001	9.938610e-001	80853
show →photograph	1.165590e+002	3.959215e-001	7.780500e+004	8.659594e-001	8.416785e-001	78044

Table 8. Comparison of precision of similarity measure by Euclidean distance and other state-of the-art approacher

Tag Pair	Euclidean Distance Between Tags	Cosine	Jaccard	Dice	Projection	Mutual information
flick→video	2.40795e+002	9.438033e-001	25182	4.026363e-001	6.976565e-001	42134
flick →picture show	2.079399e+002	8.579639e-001	4.352900e+004	5.184922e-001	6.539791e-001	2570
flick →big screen	1.551193e+002	6.065641e-001	4.333000e+004	4.674954e-001	9.227854e-001	44185
flick →silver screen	3.080568e+002	7.986752e-001	4.492700e+004	4.438021e-001	9.069607e-001	43529
flick →picture	1.414953e+002	6.228331e-001	4.213400e+004	5.218116e-001	8.316041e-001	43330
flick →photograph	1.499433e+002	6.382175e-001	4.552800e+004	5.291944e-001	8.031590e-001	32087

5. CONCLUSION

The analysis of pure dictionary based and mathematical methods encourages use of combination of these two. Dictionaries do not give complete coverage to folksonomy vocabulary. In contrast, with use of pure mathematical techniques, we can't get precise similarly for all synonyms. These two facts encourage us to opt for hybrid solution. Our proposed approach exploits both dictionaries and mathematical formula to find synonyms with accuracy. We concluded that the performance of hybrid approach is better than that of pure dictionaries and mathematical methods. In addition, Euclidean distance measures similarity more precisely as compared to Cosine similarity, Jaccard, Dice, Projection and Mutual information. In future work, we are planning to solve ambiguity problem caused due to polysemy in tagset.

ACKNOWLEDGEMENTS

The work is extended version of article published in the conference proceedings of 3rd Multi Disciplinary Student Research Conference (MDSRC – 2017). The work is part of MSc Computer Science Thesis of first three authors.

REFFERENCES

[1] F. Jabeen, S. Khusro, A. Majid, and A. Rauf. 2016. Semantics discovery in social tagging systems, A review, *Multimed. Tools Appl*,75(1), pp 573–605 .

[2] S. Hayman and N. Lothian. 2007. Taxonomy Directed Folksonomies Integrating User Tagging And Controlled Vocabularies For Australian Education Networks, in *New Developments in Social Bookmarking, Ark Group Conference: Developing and Improving Classification Schemes.*, pp. 1–27.

[3] D. Lin, S. Zhao, and H. D. District, "Identifying synonyms among distributionally similar words," 3(4), pp. 1492–1493, 2003.

[4] H. Wu , M. Zhou. 2003. Optimizing Synonym Extraction Using Monolingual And Bilingual Resources, *Proceedings of the 2th international workshop on Paraphrasing. Association for Computational Linguistics*, July 11 - 11, pp 72–79.

[5] V. D. Blondel and P. Senellart. 2002. Automatic extraction of synonyms in a dictionary, *SIAM Int. Conf. data Min.*, pp 1–7.

[6] R. Barzilay and K. R. McKeown. 2001. Extracting paraphrases from a parallel corpus, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, July 06 - 11, pp 50–57.

[7] J. R. Curran. 2002. Ensemble methods for automatic thesaurus extraction, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, July, pp. 222–229,.

[8] S. Lee and H. Yong. 2007.TagPlus : A Retrieval System using Synonym Tag in Folksonomy,pp 294-298.

[9] D. Freitag *et al.*2005. New Experiments In Distributional Representations Of Synonymy, *Proceedings of the 9th Conference on Computational Natural Language Learning. Association for Computational Linguistics*, June 29 - 30, pp 25–32.

[10] M. Piasecki, S. Szpakowicz, and B. Broda. 2007. Extended Similarity Test For The Evaluation Of Semantic Similarity Functions, *Proceedings of the 3rd Language and Technology Conference. Poznań, Poland: Poznań, Wydawnictwo Poznańskie Sp. z oo*, pp 104–108.

[11] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, and M. Pas, "A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches," no. June, pp. 19–27, 2009.

[12] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. 2009. Evaluating Similarity Measures For Emergent Semantics of Social Tagging, *Proceedings of the WWW' 09 18th international conference on World Wide Web*, April 20 - 24, pp 641–650.

[13] H. Mousselly-sergieh *et al.2013*.Tag Similarity in Folksonomies, in *Proceedings of the XXXI INFORSID congress,May 29,pp 319-334.*

[14] T. K. Landauer and S. T. Dumais . 1997.A solution to plato ' s Problem : the latent semantic analysis theory of acquisition , induction , and representation of knowledge,1(2), pp 211–240.

[15] P.Turney. 2001. Mining the Web for Synonyms : PMI-IR Versus LSA on TOEFL ,Proceedings of the 12th *European Conference on Machine Learning,Freiburg,Germany, September 5-7*, pp 491–502.

[16] E. Terra and C. L. A. Clarke. 2003. Frequency Estimates For Statistical Word Similarity Measures, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, May 27 - June 01, pp 165–172.

[17] G. Solskinnsbakk and J. A. Gulla. 2011. Mining Tag Similarity In Folksonomies, *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, October 28 - 28, pp 53–60.

[18] G. Quattrone, E. Ferrara, P. De Meo, and L. Capra. 2011. Measuring Similarity In Large-Scale Folksonomies, *Proceedings of the 23rd International Conference on Software Engineering and Knowledge Engineering* ,25 jul, pp 385–391.

[19] T. Mikolov, G. Corrado, K. Chen, and J. Dean. 2013.Vector Space,P 1–12.

[20] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. 2008. Semantic analysis of tag similarity measures in collaborative tagging systems,*Data Eng.*, 805(14), pp 1–5.

[21] A. Rêgo12, L. Marinho, and C. E. Pires. 2012.Learning Synonym Relations From Folksonomies,*IADIS Int. Conf. WWW/Internet*, pp 273–280.

[22] D. Eynard, L. Mazzola, and A. Dattolo, 2013.Exploiting tag similarities to discover synonyms and homonyms in folksonomies, vol.43,no.12,pp.1437–1457.