Lifetree: Building and Comparison based on User's Tweets

Seyedmahmoud Talebi
Dept. of Computer Science, University of Mysore,
Mysuru, India

Manoj K.
Bengaluru, India

G. Hemantha Kumar
Dept. of Computer Science, University of Mysore,
Mysuru, India

Nima Nosrati
Mysuru, India

ABSTRACT
In this paper, we create an information tree pertaining to the natural user's communication in the real world to ascertain the user's interests. This is performed by analysing users' twitter posts or tweets and comparing them with Wikipedia to generate a graph tree, with nodes pertaining to topics matched. The generated Lifetree is dynamic in nature and is progressed as the continuing users' communication i.e. is appended to the Lifetree. The various uses of the Lifetree included an overall picture of particular users' interests and further helps in event allocation, ads customization, etc...

Hence, a novel approach for representing users' data has been proposed, which makes the process of recommendation easier and more accurate. To achieve this, knowledge base and machine learning algorithms have been proposed and utilized.

Keywords
Social network, Big Data, Keyword extraction, Knowledge base, Graph analysis.

1. INTRODUCTION
Content personalization based on social activities (clicks, posts) is gaining increasing traction with web companies day by day. A variety of services and platforms on the digital web, right from movies on Netflix to navigation routes on GPS (Waze) are personalized based on what you like and what you did.

The personalized content for each individual is determined using various metrics such as click behavior, collaborative filtering and cookies. A common element across these techniques is the focus on using current browsing session for providing personalization and therefore a lack of identification of the broader interests[1].

Nowadays, a lot of users spend time on social network and share their daily activities, opinions and interest. Working with this huge number of unstructured data for any analysis such as recommendation could be difficult and exhaustive, therefore building structure which is summarization or model of user data could be helpful.

The proposed structure is labelled as 'Lifetree' in this work. Lifetree could help systems to suggest users more accurate and more relevant information that is required. As mentioned before these suggestions are such as friends, events, entertainment, shopping, etc.

To build Lifetree, we need data from user therefore social network could cover it and amongst all social networks we chose Twitter as a best source of our work with more than atleast 330 million active users (as in 2018). To build Lifetree, we leverage knowledge-base for better understanding of users' tweets. Among all knowledge-bases, Wikipedia has been chosen because it covers most of the topics and is open source.

Wikipedia stored all data in XML format. In this work, we have extracted two different graph-structured data from Wikipedia: Wikipedia Article Graph and Wikipedia Category Graph which will be explained in following sections. Wikipedia is a free multi-lingual online encyclopedia that is constructed in a collaborative effort of voluntary contributors and still grows exponentially.

During this process, Wikipedia has probably become the largest collection of freely available knowledge. A part of this knowledge is encoded in the network structure of Wikipedia pages. In particular, Wikipedia articles form a network of semantically related terms, while the categories are organized in a taxonomy-like structure called Wikipedia Category Graph [2].

Categories are tags that connected to each other therefore; it would be in graph structure. In Wikipedia all pages referring to category and these Categories has a center called "Category:Main topic classifications". It has 22 children and each one of them has variety of sub_children to cover category of any topic mentioned in Wikipedia.

As it mentioned Lifetree could be used in Social Network for people recommendation, etc. In this work, we measure the similarity between users by 1) users' Keywords and 2) users' Lifetrees

In calculation based on keywords, only keywords of the users will be calculated. However, in calculation based on Lifetree, the similarity between nodes and edges of two hierarchical tree will be calculated.

Twitter is an online social networking and microblogging service that enables its users to send and read text-based posts of up to 140 characters, known as "tweets". It was created in March 2006 by Jac k Dorsey and launched in July the same year. The service rapidly gained worldwide popularity, with over 140 million active users as of 2012, generating over 340 million tweets daily and handling over 1.6 billion search queries per day.

Twitter is primarily an "interest" based social network. Users either join Twitter to speak about things that they are interested in, or to listen to others who are talking on topics which they are interested in. Tweets are publicly visible to everyone on the web by default; however, senders can restrict message delivery to just their followers by making their accounts private. Users can tweet via the Twitter website, compatible external applications (such as Tweetdeck or

Echofon), or by Short Message Service (SMS). Users may subscribe to other users' tweets, which is known as "following" and the subscribers are known as followers. Following someone implies that all of his/her tweets will be visible on your personal Twitter homepage also known as home time-line.

Today, Twitter contains numerous celebrities, politicians, sports-person, news-media outlets, bloggers, organizations and experts on a wide array of topics. There are even more users who just use Twitter as a medium to follow these "popular" users. Since inception, Twitter has been used for a variety of purposes in many industries and scenarios. Most notable examples include 2010-11 Tunisian protests, 2011 Egyptian revolution and the 2011 Japanese Earthquake.

As a result, Twitter has become a fertile playground for various measurement and analysis studies. There are several interesting challenges and problems which have come up over the past few years, and this thesis tries to solve one of them.

Of all the knowledge bases, Wikipedia has so far, proven to be one of the most valuable resource; in fact knowledge bases such as DBPedia[3] and YAGO [4] have been derived from Wikipedia. It is an online, collaboratively generated encyclopedia and one of the largest and most consulted reference works in existence. Wikipedia is written with the goal of human consumption but it contains a certain structure, which can be exploited by automated algorithms. This structure is composed of hierarchical categories, and these categories act as semantic tags to different Wikipedia articles. Moreover, each article interlinks with each other using the anchor text within the content. Recent years have seen many significant research questions being solved with the help of Wikipedia[5], [6], [7], [8] and it has been successfully applied to complement the understanding of different datasets[9].

2. LITERATURE SURVEY

Recently, the ability to discern topics from social media has begun to receive attention as the necessity to search the text and user profiles gains importance. In fact, two recent approaches address different problems using Twitter data.

In [10] explore the problem of recommending content (Tweets). They build a number of recommender approaches, one of which is "topic" based. They model the topics of a user as a bag-of-words generated from the user's Tweets (with TF/IDF weights).

They then compare this feature vector modelling of the topics to a similar feature vector of an incoming Tweet to determine if it should be recommended to the user. There are a few drawbacks to this approach. First, because a bag-of-words is used, the terms must be very specific.

Another approach to analyzing Twitter that uses topics is TwitterRank, which aims to identify influential micro-bloggers [11]. This approach leverages LDA by creating a single document from all of a user's Tweets and then discovering the topics by running LDA over this "document." Again, such an approach has the problems of LDA since the Twitter data is sparse, and the generated topics are based on terms rather than concepts.

There is also work that is similar in that the goal is to determine the topics and interests of bloggers by analyzing their blogs[12],[13],[14]. However, blogs are a much richer medium for textual analysis because blog posts are generally much longer than Tweets and usually conform better to the grammatical rules of written English. In the area of web personalization and recommendation, generating hierarchical interests for a user involves analyzing web documents. In [15] the authors have realized top-down techniques to hierarchically cluster web documents the user is interested in. Both the techniques are built upon Bag Of Words approach and the hierarchical clusters of terms form the user profiles. On the other hand, work in [16],[17] analyze web documents and leverage ontologies to create contextual user profiles. The former [16] use Bag Of Words approach to map web documents to Wikipedia concepts. [17] used DMoz with an adaptation of spreading activation to map web documents to DMoz articles.

User interests extracted from social messages have been represented as Bag Of Concepts in various works [18], [19], [20]. One of the main aspects of these works is the weighting schemes used to reflect user's interests towards the concepts. Abel et al. in their work [18] compare hashtag-based, entity-based and topic-based user models generated from tweets, for news recommendation. The approach scores the concepts/interests based on simple term frequency technique. The same technique is employed by TUMS system developed by Tao et al. [20]to generate semantic user profiles, provides an aggregated score for concepts from multiple social networks (Facebook and Twitter) with a temporal decay. Other techniques such as tf-idf, temporal scoring [21] have also been used to score interests. Although, it will be interesting to evaluate the impact of these scoring mechanisms on the weights of interest categories in HIG.

Wikipedia Graph has been leveraged as the base for generating HIG in the approach. Other approaches have utilized it for tasks such as ontology alignment [22], and clustering [23].Further, Spreading Activation theory used in our approach to assign interest scores has also been adapted to tasks such as document categorization [24] and search results personalization [17].

3. PROPOSED METHOD

The goal of this approach is to construct "Lifetree", which would represent user lifestyle and interest. In the second part, similarity is measured between users based on two different methods. The inputs for this approach are:

1) Tweets of user, which is the main source for this system to figure out her/his interest
2) Category graph extracted from Wikipedia
3) Articles of Wikipedia which has been named wiki_converted in this work.

3.1 Building Lifetree

First, we get users' tweets, after some pre-processing, keywords will be extracted. Based on the keywords, proper pages of Wikipedia will be selected. Each page of Wikipedia contains list of categories which the page belongs to. After finding the proper pages and the categories of that, in the next step from each category which has been selected to the root of Wikipedia Category Graph, shortest path would be chosen and the whole path would be added to Lifetree as a new branch of hierarchical interest of user. At the end of the process we would have hierarchy tree with multiple level of categories which from top level it mentions generic topics of user and depends on the Wikipedia classification the sub-categories will be listed.

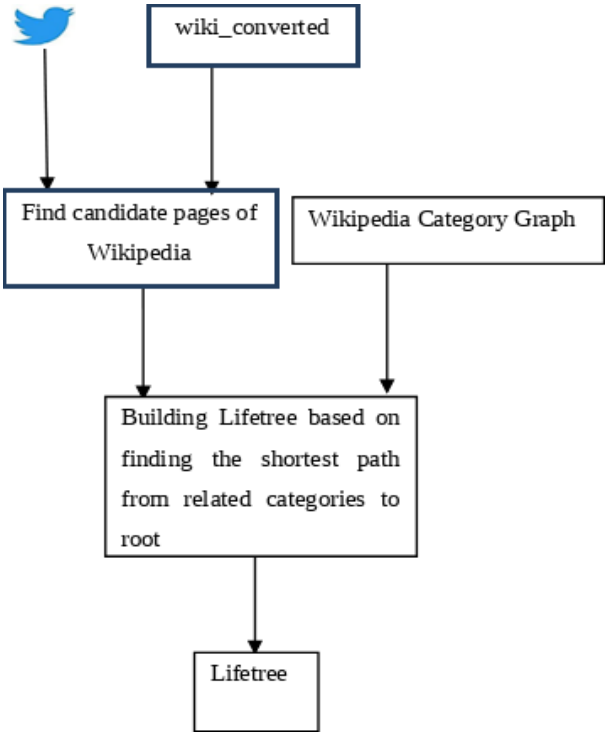

Fig 1: Overall Steps for Building Lifetree

The whole Wikipedia will convert to two different file for faster analysis. The first one is wiki_convert that is only the article pages (Wikipedia Article Pages). In wiki_converted file in page tag title of page and categories which the page connected has been mentioned. This file is extraction of Wikipedia which help us better and faster process and the categories of those and the second one is Wikipedia Category Graph that is a graph type of data covers whole categories and connection between them.

Figure 1 shows the flowchart of the process. After preparing these two source of data, the whole process will be as below.

Step1: Pre-processing such as removing unwanted words from tweets and removing duplicate posts.

Step2: Extracting keywords with RAKE method

Step3: Find candidate pages related to keywords from wiki_converted

Step4: List all the categories, which candidate pages are related to.

Step5: For each category in step4 find shortest path to WCG and add the path to Lifetree. Therefore, Lifetree has root "Category: Main Topic Classification" and all categories which are related to keywords listed in hierarchical structure.

3.2 Comparing similarity between users

All user in social network are interested in to make new friendship or join pages of sports, entertainment or events therefore suggesting more related to user interest would be more amazing and useful. In this work by comparing, the model that has been extracted from twitter users will be used for measuring similarity.

In this work, to calculate the similarity two different approaches have been used, which are based on:

1) Keywords (using cosine similarity)
2) Lifetree (custom matching of nodes or edges)

3.3 Experimental Analysis

The whole process needs two inputs, user tweets and Wikipedia. We have used API to retrieve users' tweets. Five different users in different categories has been selected and we try to extract Lifetree and calculate similarity between them.

As discussed, we need Wikipedia to provide knowledge to the system. Wikipedia dump 2017 has been used which is 63.3 GB file in XML format. As mentioned to make process faster two different format of Wikipedia has been extracted *wiki_converted* and *wiki_category*. wiki_converted contains only Wikipedia article pages we removed namespaces. In wiki_category only pages which contains "Category:" in the title of the page and subcategories of them will be extracted.

Table 1. Experimental results of Tom Hanks

Posts	Number of Keywords	Matched pages	Number of categories
0-50	16	8	52
50-100	15	20	62
100-150	14	8	58
150-200	25	8	44

Table 2. Experimental results of Bill Gates

Posts	Number of Keywords	Matched pages	Number of categories
0-50	56	11	71
50-100	61	3	22
100-150	58	1	12
150-200	50	5	56

Table 3. Experimental results of NASA

Posts	Number of Keywords	Matched pages	Number of categories
0-50	51	4	58
50-100	54	4	50
100-150	57	3	12
150-200	56	5	54

In tables 1, 2, 3 we got various statistics of different users. We chose the first fifty, second fifty etc, tweets of the users. In the tables 1, 2, 3 "matched pages" are the number of pages in wiki_converted are relevant to keywords extracted for each list of tweets and "number of categories" are the collection of categories each candidate page contains. For example, page1, page2 …, page8 are the candidate pages of "wiki_converted",

each of this candidate pages have categories which they connected to, therefore the collection of all categories will be "number of categories".

We can observe from above tables more number of keywords is not a reason for more number of categories matching, because not all keywords are matched in Wikipedia.

The categories listed for each analysis are one node of "wiki_category" graph. For each of these categories, we select the shortest path to root of the Wikipedia Category Graph. Moreover, each of these paths will become branches of Lifetree.

For better understanding of the output, Lifetree of user Bill Gates and Leonardo DiCaprio has been represented in Figures 2,3. It is radial representation of the graph. Center Root of the graph, which has blue color, contains the name of the twitter user.

Figure 2 is the Lifetree of Bill Gates who is mostly in technology and business (and of course, other fields as well). If we consider the root of the tree, which is ROOT_BILLGATES as a first level, we can say that in second level this user interests covered 12 major topics out of 22. These 12 topics are such as *["Games","Politics","Health",…,"Mathematics"]*. We numbered 22 because it is the number of topics Wikipedia considered for the coverage of all the concepts exist. Depth of the tree is varied from 3 to 6 depends to the situation and distance of the categories to the root.

In Figure 3, which is Dicaprio's Lifetree the most topic discussed are in the field of music, arts, technology and society.

After Building Lifetree, the next process is to calculate the similarity between users. As mentioned previously we used two different methods of calculation. Measurement based on only keywords of user tweets and based on Lifetree. In similarity based on keywords, cosine similarity method will be considered. Table 4 is the result of users based on cosine similarity. It shows the similarity between two users. For example, similarity between twitter users: NASA and Tom Hanks is 24%.

Table 4. Similarity between Users based on Cosine Similarity Methodology

Twitter User	Bill Gates	NASA	Tom Hanks	Elon Musk	Leonardo DiCaprio
Bill Gates	1	0.35	0.38	0.37	0.46
NASA	0.35	1	0.24	0.31	0.29
Tom Hanks	0.38	0.24	1	0.31	0.20
Elon Musk	0.37	0.31	0.31	1	0.22
Leonardo DiCaprio	0.46	0.29	0.20	0.22	1

These kind of measurements are very fast and using fewer resources, specially memory but they just simply without having more data and details of user try to find similarity, which could not be appropriate, therefore using Lifetree could be a good resource for the purpose of similarity measurement.

After building Lifetree, the method used for distance measurement called Custom Distance Measurement of Lifetree (CDML). Here is how this method works:

Step1: Load XML graph Lifetree; Load XML graph Lifetree2.

Step2: Find number of edges common in both Lifetrees. For each edge in Lifetree1, check if present in Lifetree2 and if exists then is added to common list.

Step3: Calculate the length of common edges divide by number of both Lifetrees.

Initially, we load the files which is kind of text file format. Which first lists the entire node name and their given id, then list nodes pair between edges. Below is a sample of such kind of file:

1. Main topic classifications
2. Geography
3. Places
4. Astronomical objects
5. Black holes
6. Humanités
7. Fiction
8. Galaxies in fiction
9. Galaxies
10. Mathematics
11. Applied mathematics
12. Mathematical physics
13. Theory of relativity
14. Articles

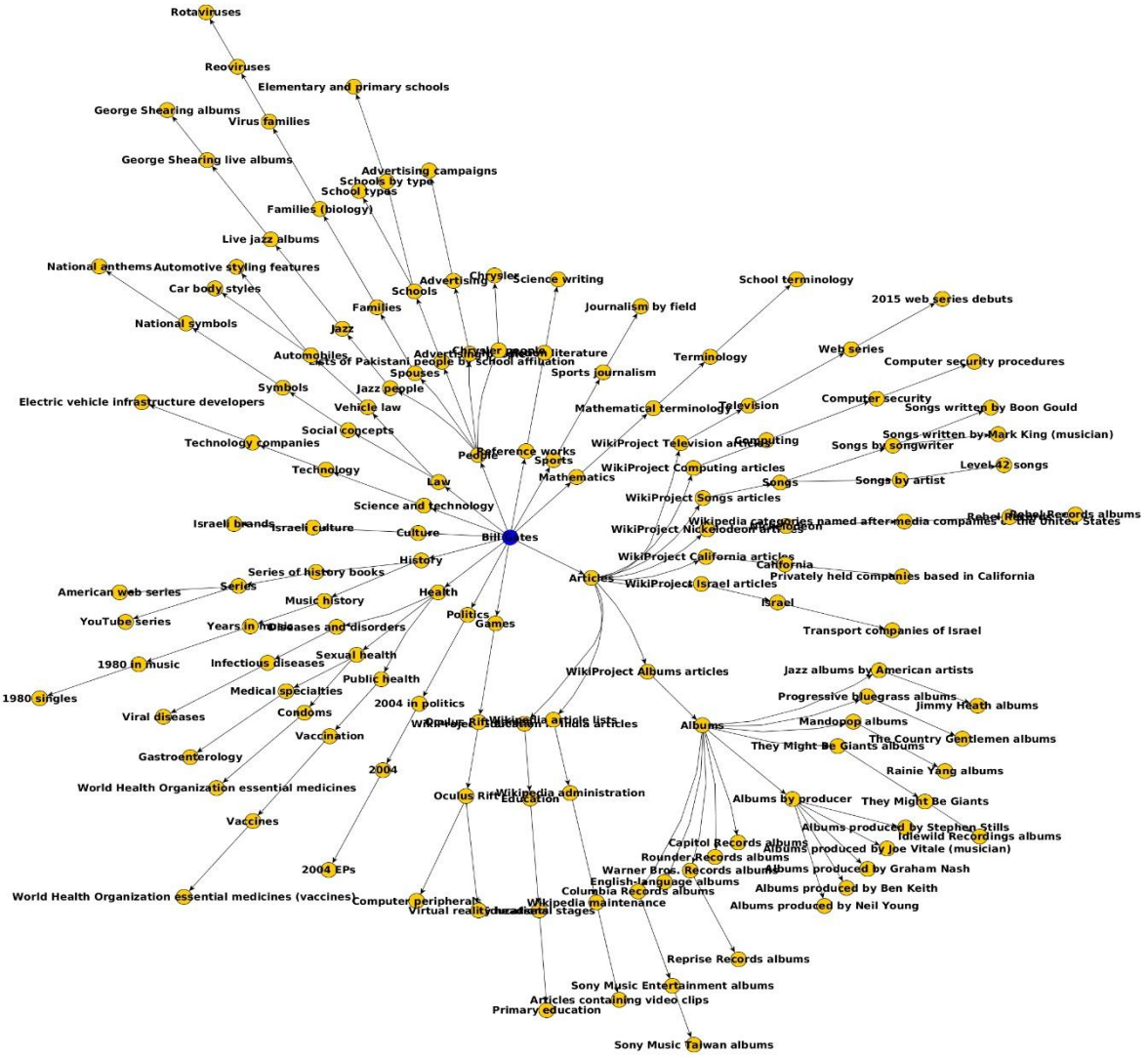

Fig 2. Bill Gate's Lifetree

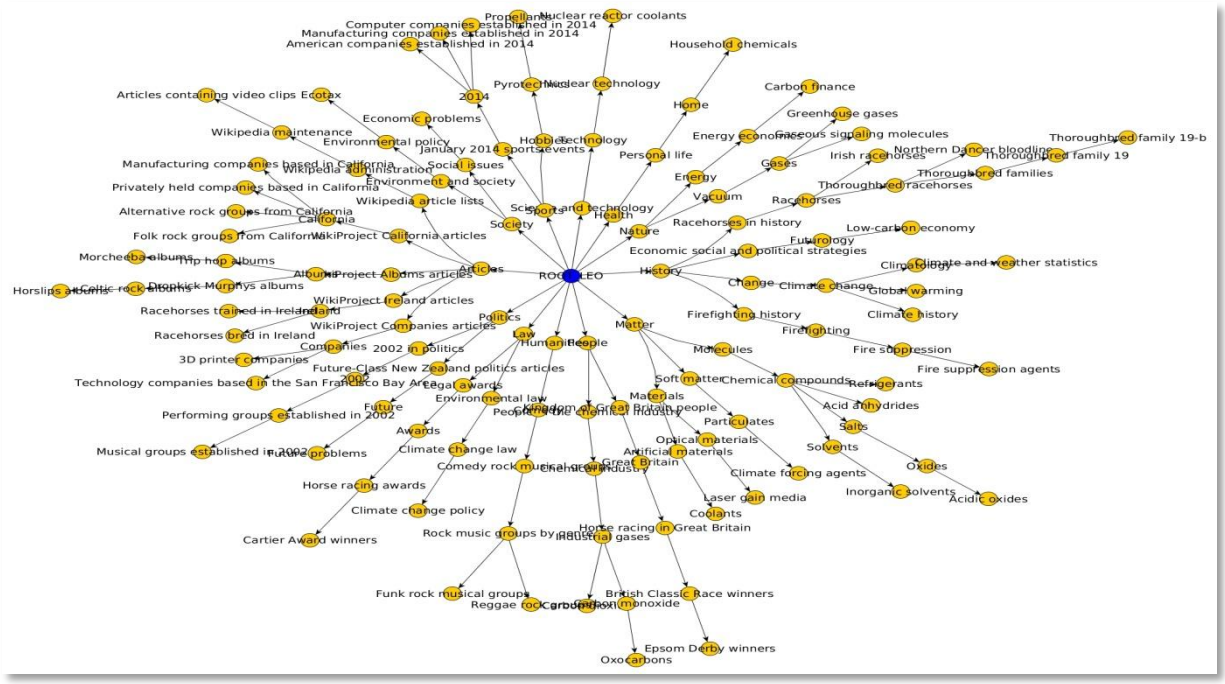

Fig 3. Leonardo DiCaprio's Lifetree

These files divided to two part which separated by "#", the first part is list of the nodes name and their ids which would be refer in second part. In second part it is list of edges between two nodes which start from root or on the other hand node id=1 till all the branches of the Lifetree.

In the next steps, we find common edges between Lifetrees and can keep it in separate file. These three files are enough to help us for measuring similarity by calculating the percentage of common edges between two Lifetrees. We used these measurements for five users to do more analysis and comparison with keyword similarities.

Table 5. Similarity between Users based on their Lifetree's

Twitter User	Bill Gates	NASA	Tom Hanks	Elon Musk	Leonardo DiCaprio
Bill Gates	1	0.11	0.10	0.10	0.07
NASA	0.11	1	0.07	0.14	0.04
Tom Hanks	0.10	0.07	1	0.07	0.04
Elon Musk	0.10	0.14	0.07	1	0.05
Leonardo DiCaprio	0.07	0.04	0.04	0.05	1

Refer to Tables 4 and 5, in calculating based on cosine similarity for Bill Gates, the minimum similarity is NASA with 35% and maximum similarity is Leonardo DiCaprio with 46% while in measurement based on Lifetree it is vice versa. For better understanding, we considered NASA and compared it in both results. For NASA the best match is Bill Gates with 35% similarity and the minimum is Tom Hanks with 24% similarity while in Lifetree matching, maximum is with Bill Gates with 11% and minimum is Leonardo DiCaprio 4% which could be reasonable for our own understanding

In table 4, the maximum similarity among all users is between Bill Gates and Leonardo DiCaprio with 46 % similarity and minimum is between Tom Hanks and Leonardo DiCaprio with 20% similarity.

In table 5, which is based on Lifetree matching the strongest connection is between NASA and Elon Musk with 14% similarity and minimum connection is between NASA and Leonardo DiCaprio and Tom Hanks and Leonardo DiCaprio with 4% similarity.

4. CONCLUSION

With Cosine Similarity all the keywords are compared between users, but in the customized function exact matching of the Lifetree's nodes are performed. Therefore, the similarity obtained is stronger and less random in the node matching and is recommended.

Lifetree represents the culmination of users' interest as a continued progression and gathering through social media posts, in our case: Twitter. The more user generates new posts on twitter; the Lifetree of that user is dynamically analyzed by adding new nodes. While adding new nodes in the Lifetree, we ensure that there is minimal duplication. In addition, we can incorporate newer conditions or methods while adding information to the Lifetree, thereby making the proposed structure to be scalable.

Lifetree helps to predict users' interests and is extremely useful for targeted advisements, among other uses such as behavioral analysis and predictions, which could be extremely useful in artificial intelligence systems.

5. REFERENCES

[1] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User interests identification on twitter using a hierarchical knowledge base," in European Semantic Web Conference, 2014, pp. 99–113.

[2] T. Zesch and I. Gurevych, "Analysis of the Wikipedia category graph for NLP applications," in *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, 2007, pp. 1–8.

[3] C. Bizer *et al.*, "DBpedia-A crystallization point for the Web of Data," *Web Semantics: science, services and agents on the world wide web*, vol. 7, no. 3, pp. 154–165, 2009.

[4] G. Kasneci, F. Suchanek, and G. Weikum, "Yago-a core of semantic knowledge," 2006.

[5] P. Ferragina and U. Scaiella, "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1625–1628.

[6] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *IJcAI*, 2007, vol. 7, pp. 1606–1611.

[7] E. Meij, W. Weerkamp, and M. De Rijke, "Adding semantics to microblog posts," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 563–572.

[8] I. H. Witten and D. N. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," 2008.

[9] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[10] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: experiments on recommending content from information streams," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1185–1194.

[11] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270.

[12] M. Oka, H. Abe, and K. Kato, "Extracting topics from weblogs through frequency segments," in *Proc. of the Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[13] C.-Y. Teng and H.-H. Chen, "Detection of bloggers'

interests: using textual, temporal, and interactive features," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, pp. 366–369.

[14] Y. Cheng *et al.*, "Model bloggers' interests based on forgetting mechanism," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1129–1130.

[15] D. Godoy and A. Amandi, "Modeling user interests by conceptual clustering," *Information Systems*, vol. 31, no. 4, pp. 247–265, 2006.

[16] K. Ramanathan and K. Kapoor, "Creating user profiles using wikipedia," *Conceptual Modeling-ER 2009*, pp. 415–427, 2009.

[17] A. Sieg, B. Mobasher, and R. Burke, "Web search personalization with ontological user profiles," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 525–534.

[18] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in *Extended Semantic Web Conference*, 2011, pp. 375–389.

[19] P. Kapanipathi, F. Orlandi, A. P. Sheth, and A. Passant, "Personalized filtering of the twitter stream," 2011.

[20] F. Orlandi, J. Breslin, and A. Passant, "Aggregated, interoperable and multi-domain user profiles for the social web," in *Proceedings of the 8th International Conference on Semantic Systems*, 2012, pp. 41–48.

[21] M. Albakour, C. Macdonald, and I. Ounis, "On sparsity and drift for effective real-time filtering in microblogs," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 419–428.

[22] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh, "Ontology alignment for linked open data," in *International Semantic Web Conference*, 2010, pp. 402–417.

[23] T. Xu and D. W. Oard, "Wikipedia-based topic clustering for microblogs," *Proceedings of the Association for Information Science and Technology*, vol. 48, no. 1, pp. 1–10, 2011.

[24] P. Schönhofen, "Identifying document topics using the Wikipedia category network," *Web Intelligence and Agent Systems: An International Journal*, vol. 7, no. 2, pp. 195–207, 2009.