INTELLIGENT SOCIAL NETWORKS MODEL BASED ON SEMANTIC TAG RANKING

Rushdi Hamamerh [1] and SamehAwad [2]

[1]Department of Computer Engineering, Al-Quds University, Abu Dies, Palestine
[2]Department of Information Technology, Birzeit University, Ramallah, Palestine

ABSTRACT

Social Networks has become one of the most popular platforms to allow users to communicate, and share their interests without being at the same geographical location. With the great and rapid growth of Social Media sites such as Facebook, LinkedIn, Twitter…etc. causes huge amount of user-generated content. Thus, the improvement in the information quality and integrity becomes a great challenge to all social media sites, which allows users to get the desired content or be linked to the best link relation using improved search / link technique. So introducing semantics to social networks will widen up the representation of the social networks.

In this paper, a new model of social networks based on semantic tag ranking is introduced. This model is based on the concept of multi-agent systems. In this proposed model the representation of social links will be extended by the semantic relationships found in the vocabularies which are known as (tags) in most of social networks.The proposed model for the social media engine is based on enhanced Latent Dirichlet Allocation(E-LDA) as a semantic indexing algorithm, combined with Tag Rank as social network ranking algorithm. The improvements on (E-LDA) phase is done by optimizing (LDA) algorithm using the optimal parameters. Then a filter is introduced to enhance the final indexing output. In ranking phase, using Tag Rank based on the indexing phase has improved the output of the ranking. Simulation results of the proposed model have shown improvements in indexing and ranking output.

KEYWORDS

SocialNetwork, Multi-Agent Systems, Semantic Indexing, Tag Rank, LDA, E-LDA.

1. INTRODUCTION

Social networks are emerging field in information interchange, worldwide used and wanted. It is a challenging subject to do a research in social media field as it was and still affecting us in every aspect of our lives [1].

Ellison and Boyd defined social networks (SN) as web-based services that allow users to build a public or semi-public profile within a system, connect to a list of other users by sharing a connection, and view and extend their list of connections and those made by others within the system. The nature of these connections may vary from (SN) site to another [2].

In current social networks, Links between contents are constructed by many ranking techniques according to the way to deal with data, importance and priority of data. Such as posts in Facebook, hashtags in Twitter, Job and Experiences in LinkedIn, etc. and so data must be ranked in a way that links constructing the social graph will reflect natural distribution and connection between nodes of the social networks. Rank of each node is given by making iterative process of weights in network. In Semantic Social Networks, this weight can be given according to semantic content of the social network node.

Semantic Content of Social Network which is large and complex collections of data and that is known nowadays as “Big Data” [3] must be indexed before ranking process. This can be achieved

by introducing semantic indexing algorithms to process content of Social Networks [4].Improving indexing output and choosing the proper rank algorithm will affect the quality of the social graph and how nodes will be linked in social network.

2. MULTI-AGENT SYSTEMS

For Indexing and Ranking processes. The Concept of Multi-Agent system (MAS) is a great addition to give good, improving, and self-learning mechanism especially in social networks. Multi-Agent Systems are computerized system consisted of multiple agents that they interact intelligently within the environment which can be used to solve problems [5].

The agent precepts data and grabs the documents from the environment (SSN) and using the learning elements it updates the performance elements. And it builds the knowledge base by updating the learning elements based on critics that represents the feedback from the whole system that the agent is working in. and this knowledge base generates problems that can be used as condition rules to be used in the decision that the agent will make to do the needed action in the environment.

In social networks, the multi-agent implementation theories have two main perspectives: user perspective and network perspective.

In user perspective, the agents will be the user accounts [6], which means each account will act as agent in mediating data and negotiating connections with the other agents to enlarge their social network.

Nevertheless, in the network perspective, the agents will be carrying out some central operations such as filtering data, managing connectivity, and building the social graph.

Because of the semantics in social network is being discussed, the perspective of semantics must be an important role to be done by agents in the multi-agent implementation of social network.
In this paper, the concentration will be on some roles done by agents, which are parsing data, building semantic index of the data, then ranking this index, and finally build connections between contents according to the rank output.

3. SEMANTIC INDEXING - LATENT DIRICHLET ALLOCATION (LDA)

Indexing algorithms - mainly in search engines - collect, parses, analyse and store data to facilitate quick and accurate information retrieval [7]. Index design includes interdisciplinary concepts from linguistics, cognitive psychology, mathematics, computer science and informatics. An alternative name for the process in the context of search engines intended for searching web pages on the Internet is web indexing.

When dealing with information retrieval, stored documents are identified by sets of terms that are used to represent the contents of the document. The indexing process is the assignment of the index for documents in the collection of documents. The index of terms can be predefined as a fixed set of controlled vocabulary or can be any additional words that the indices consider to be related to the topic of the document.

One of the most popular indexing algorithms is Latent Dirichlet Allocation (LDA) [8], is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities.LDA assumes that each document contains different topics, and words in the document are generated from these topics. All documents contain a

specific set of topics, but the proportion of each topic in each document is different. The generative process of the LDA model can be described as follows [9]. Assuming document w in a corpus D:

1- Choose N ~ Poisson (ξ).
2- Choose θ ~ Dir(a).
3- For each of the N words w_n:

Choose a topic z_n ~ multinomial distribution (θ)
Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

Many simplifying assumptions are made in this basic model, such as removing some subsequent sections.

First, the dimensionality k of the Dirichlet distribution which means the dimensionality of the topic variable z is assumed known and fixed. Second, the word probabilities are parameterized by k $k \times V$ matrixβ where $\beta_{ij} = p(w^j = 1 \,|z^i = 1)$ which for now is treated as a fixed quantity that is to be estimated.Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed.

Furthermore, note that N is independent of all the other data generating variables (θ and z). It is thus an ancillary variable and its randomness will generally be ignored in the subsequent development. A k-dimensional Dirichlet random variable θ can take values in the (k−1)-simplex (a k-vector θ lies in the (k−1)-simplex if $\theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1$), and has the following probability density on this simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1} \quad (1)$$

Where the parameter α is a k-vector with components α_i>0, and where $\Gamma(x)$ is the Gamma function.The Dirichlet is a convenient distribution on the simplex—it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution.Given the parameters α and β, the joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (2)$$

Where $p(z_n|\theta)$ is simply θ_i for the unique i such that$z_n^i = 1$. Integrating over θ and summing over z, then the marginal distribution of a document will be:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \quad (3)$$

Finally, the probability (or the log-likelihood) of generating corpus is:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \quad (4)$$

4. PROPOSED MODEL

The proposed model for semantic tag ranking for social networkis based on enhanced LDA. The input in this model is the document collection where it contains the word per document count. Then the final output will be the ranking of the tags, which are the Tag Rank results of the topics index.The proposed model is based on two main phases: the indexing phase which is carried out by the indexing agent, and the ranking phase which is carried out by the ranking agent.In indexing phase, the input is the document collection where it contains word and document count. In this phase, the initialization is done then document parsed to get the initial index to be

processed by (LDA) algorithm. The output of this phase is the semantic index, which contains word-per-topic distribution and topic-per-document distribution.

In this proposed model, the focus on the topic-per-document to be processed as tags. Therefore, in the next phase, which is the ranking phase the input will be the topic-per-document distribution that came as index matrix. In ranking phase, the input will be processed by Tag Rank algorithm with the help. The final output will the Tag ranking matrix that will be sent to build the social links in the semantic social network. Figure1. Shows the proposed model architecture.

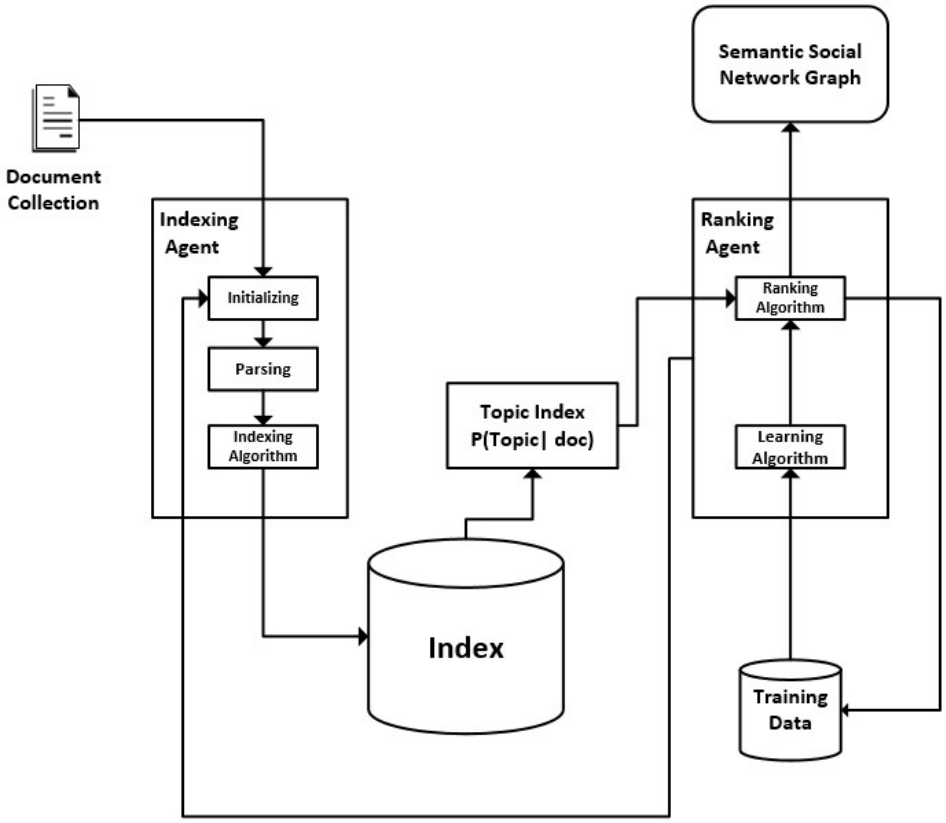

Figure1: System Architecture.

The indexing phase has seven sequential steps to build the topic per document index document based on the document collection to be processed. Figure2. Shows the steps of this phase:

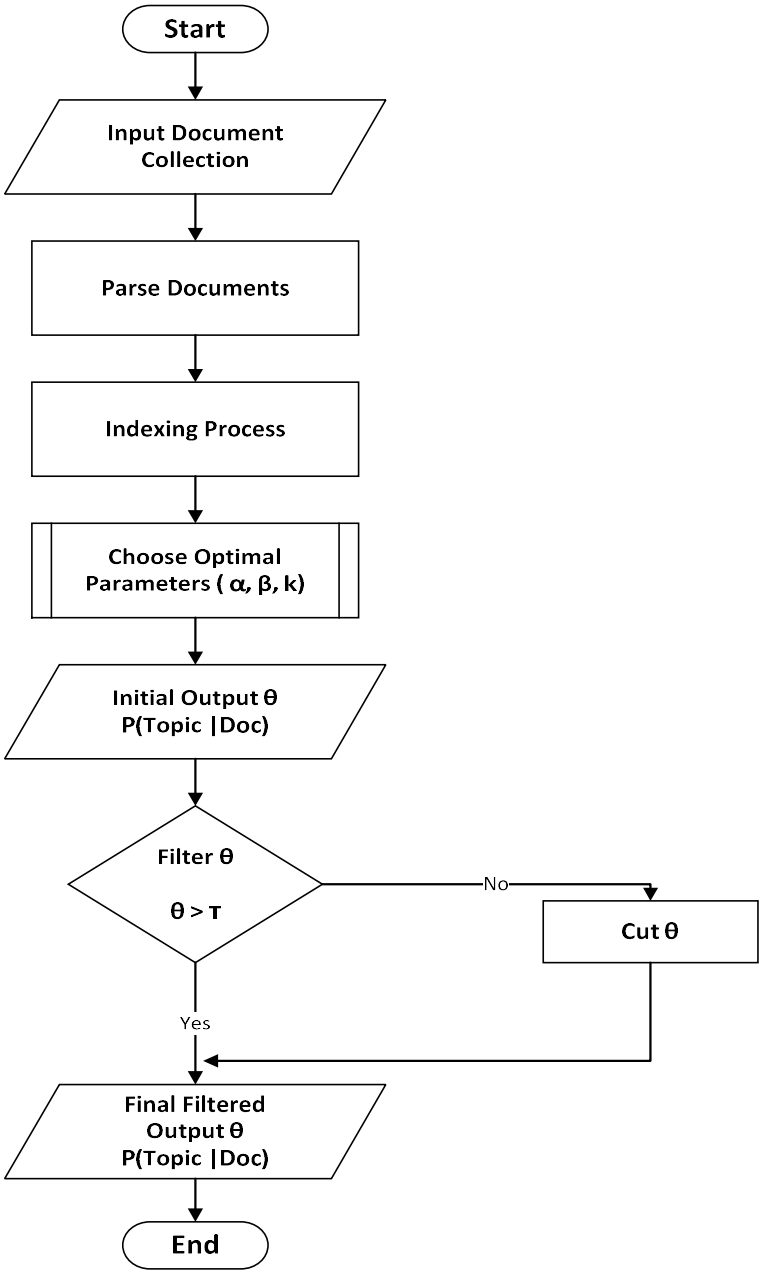

Figure2: Flowchart of Indexing Phase.

The start is with the input of document collection, which is parsed then indexed with choosing the optimal parameters ($a, \beta and K$) which increases the precision and recall of the output.

Then the output will be probability of topic per document that will be filtered by specific threshold (τ) that will be chosen by experiment in the simulation. The final output will be the filtered (θ)which is the output of the enhanced LDA algorithm. Which is called E-LDA.

The next phase is ranking phase. It starts with the output of E-LDA algorithm with checking that (θ) is higher than the threshold (τ). Then the Tag Rank algorithm start to rank (θ) as initial tag rank. The ranking algorithm is simply here to maximize the rank. Each document will get the

higher topic ranking to be the first tag. In addition, documents will be descendingranked for each tag i.e. for each topic. Figure3. Shows the steps of ranking phase.

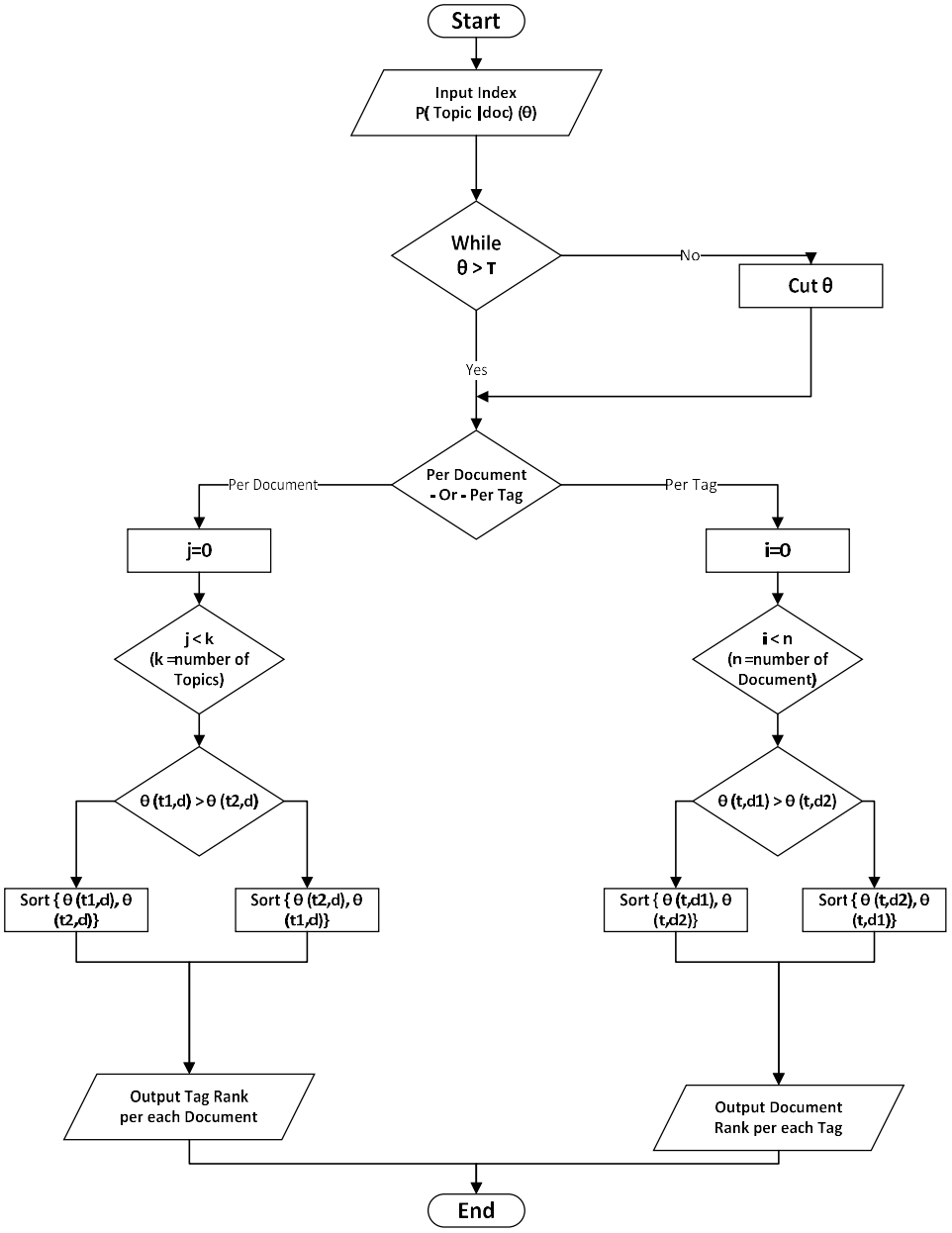

Figure3: Flowchart of Ranking Phase

As shown in these flowcharts it is obvious that there are two main intelligent agents that are carrying out the system functions. Indexing and ranking agents. The next pseudocode shows the steps of the algorithm of the semantic tag ranking.

Algorithm 1. Intelligent Semantic Tag Ranking

Input: Document Collection
Start
//Indexing Agent{
Rule 1: Get Document
Rule 2: Parse Document Content
for *i*=*1*to *n* do //n= number of document records
Rule 3: Start LDA Indexing Algorithm
end for
Rule 4: Filter
{

for *i*=*1*to *n* do //n= number of document records
Select $\theta_{t_i} where \theta_{t_i} > \tau$//$\tau$ is threshold
end for

}
Output Index ($\theta_{t_1}, \theta_{t_2}, \dots, \theta_{t_n}$)
end } //end of indexing agent job

Input: Index ($\theta_{t_1}, \theta_{t_2}, \dots, \theta_{t_n}$)
//Ranking Agent{
Start

for *i*=1 to *n* do //n= number of tags
//repeat until all tags which have larger ranks than threshold τ

Repeat{
//select document 1 and document 2 to be compared and maximized

Select Max(θ_i, θ_{i+1})

Condition: While (Max(θ_i, θ_{i+1}) $\geq \tau$) { // τ is threshold

Select Max(θ_i, θ_{i+1})

Sort (θ_i, θ_{i+1})
}
***i*=*i*+1**
} // until (all tags which are larger than τ are processed).
for *j*=1 to *k* do //k= number of documents
//repeat until all documents which have larger ranks than threshold τ
Repeat{
//select tag 1 and tag 2 which are columns and rows of Max($\theta_{t_j}, \theta_{t_{j+1}}$)
Select Max($\theta_{t_j}, \theta_{t_{j+1}}$)
Condition: While (Max($\theta_{t_j}, \theta_{t_{j+1}}$) $\geq \tau$) { //τ is threshold

Select Max($\theta_{w_j}, \theta_{w_{j+1}}$)
Sort ($\theta_{w_j}, \theta_{w_{j+1}}$)
}
***j*=*j*+1**
} // until (all tags which are larger than τ are processed).
Build Links between Tags
Output Tag Rank records
end } //end of Ranking Agent job.

5. SIMULATION RESULTS

This section presents the simulation experiment for the proposed model. The concept of combining index resulting from LDA with threshold applied to be as the Tag input for the ranking algorithm has to be proven by results and providing a good comparison between the proposed model phases, in both indexing and ranking phases

Simulation was carried out using MATLAB R2016a simulation software under Microsoft Windows 10 operating system.

The hardware platform that carried out the software is Intel core i7-3520M processor with 8 Gigabyte random access memory.

The simulation on the indexing phase will be carried out based on previous simulation works done by The Natural Language Processing Group at Stanford University [10], also on natural language labs on Iowa State University [11], and the research toolbox from University of California, Irvine [12], using their MATLAB functions to implement the enhanced LDA function. The dataset is used was *psychreview* dataset. Which contains Psychology Review Abstracts and collocation Data. This dataset contains about 85000 records of words and documents. With the initial count of words for each document and the topic.

To evaluate the simulation, main four metrics were introduced; two for evaluating indexing which are precision and recall. The other two is for evaluating ranking and these metrics are mean average precision (MAP) and Normalized Discounted Cumulative Gain (NDCG) [13].

Precision: is the ratio of the number of relevant documents retrieved to the total number of irrelevant and relevant documents retrieved:

$$Precision = \frac{|relevant\ documents \cap retreived\ documents|}{|retreived\ document|} \tag{5}$$

Recall: is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the dataset:

$$Recall = \frac{|relevant\ documents \cap retreived\ documents|}{|relevant\ document|} \tag{6}$$

Mean Average Precision (MAP): is the precision-at-k score of a ranking y, averaged over all the positions k of relevant documents:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{7}$$

Where:

$$AveP = Average\ Precison = \frac{\sum_{k=1}^{n}(P(k) \times rel\,(k))}{number\ of\ relevant\ documents} \tag{8}$$

Q is the number of queries, and:

$$rel(k) = \begin{cases} 1, & when\ item\ at\ rank\ (k) is\ relevant \\ 0 \end{cases} \tag{9}$$

Normalized Discounted Cumulative Gain (NDCG): is a normalization of the Discounted Cumulative Gain (DCG) where (DCG) is a weighted sum of the relevancy degree of the ranked items:

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{10}$$

$$\text{Where: } DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2\, i}. \tag{11}$$

And $IDCG_p$ is the ideal DCG at position p

Based on previous research done earlier [14], the chosen the optimal parameters for LDA algorithm(α,β, and k), were $k = 4,\ \alpha = \frac{0.7}{k}$, And $\beta = 0.1$.

The next enhancement is to choose the best threshold (τ) to filter the output of the indexing process. Therefore, for the index output with the parameters that have been chosen before,

calculating the precision and recall of the output with applying the filter. The result was as shown in Figure4.

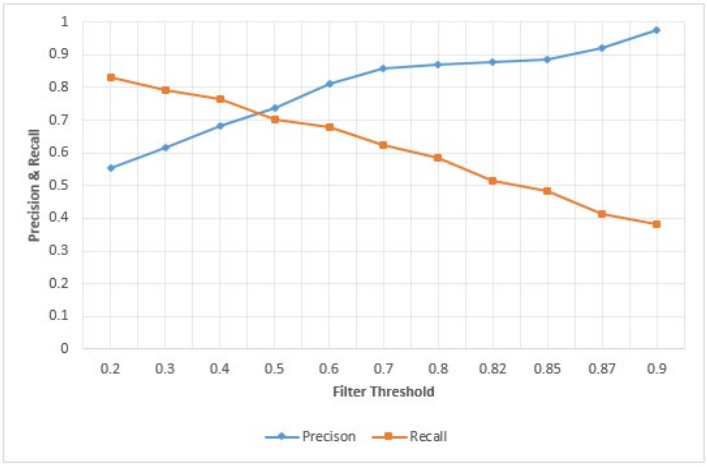

Figure 4. Precision and Recall according to (τ)

It was noticed that the best combination of precision and recall is around τ=0.5. And so it is a good suggestion to choose this value as the threshold of the filter. The resulting algorithm with these enhancement is called "Enhanced LDA" abbreviated (E-LDA). Figure5. Shows how topic distribution is enhanced using this filter:

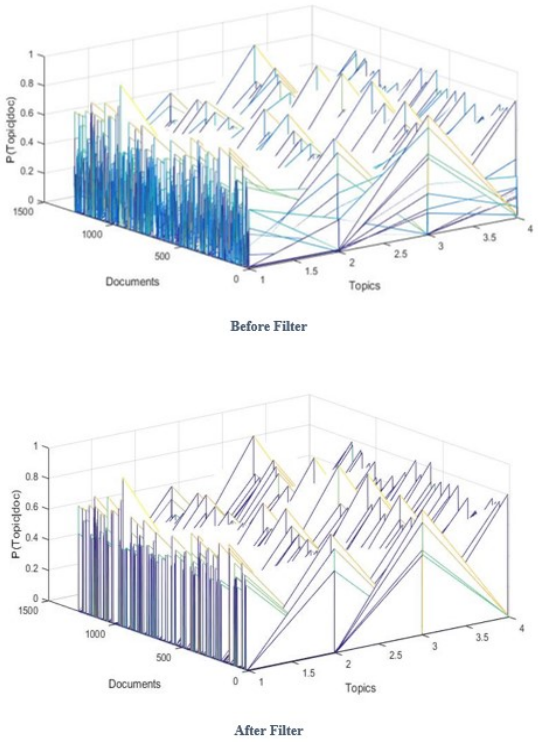

Figure5. Topic Distribution in Document Collection according to the Filter.

The simulation for the indexing agent was carried out based on previous researches comparing indexing algorithms [9] [15]. Figure6 shows a comparison between (E-LDA) and these algorithms.

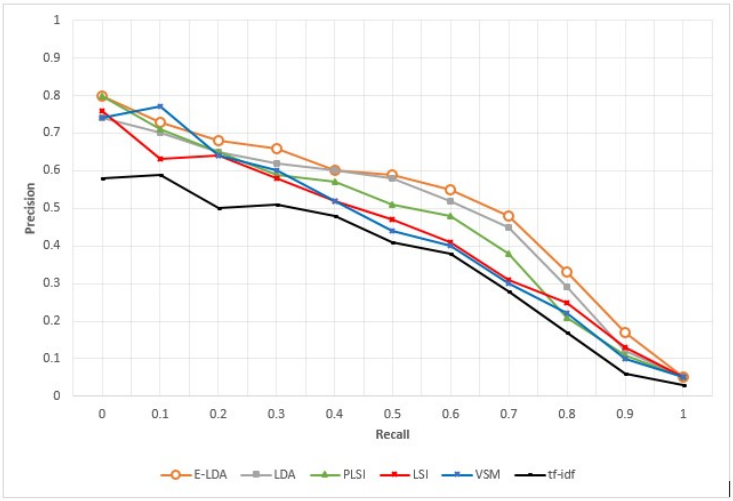

Figure 6. (E-LDA) vs semantic indexing algorithms

As shown in Figure. E-LDA is enhanced from LDA with (4%). E-LDA has better precision vs. recall combination which means better relevancy in index output.

After indexing phase enhancement done,it is possible to combine tag rank with the output to get the semantic tag rank. Figure7. Shows the improvement in precision and recall between E-LDA and the Tag Rank.

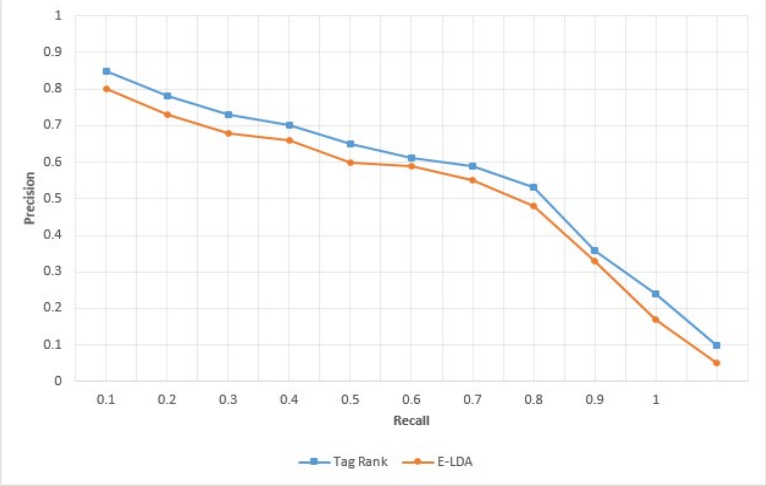

Figure 7.Tag Rank vs. (E-LDA).

As shown in Figure7. Tag Rank shows better precision and recall than input from E-LDA with almost (5%).

Comparing Tag Rank with Page Rank (PR), Weighted Page Rank (WPR), Hyper-link Induced Topic Search (HITS) and Time Rank (TSPR) [16] [17] [18] [19]. And according to MAP and NDCG@(k=4) as (k=4) is the best parameter for the indexing algorithm LDA that was concluded earlier [14]. Figure8. Shows the comparison between ranking algorithms:

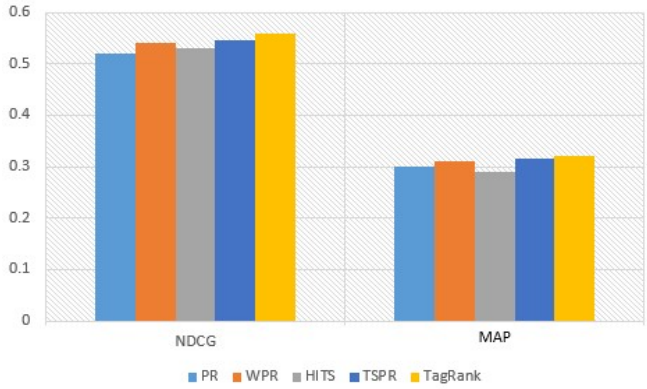

Figure 8. Semantic Tag Rank vs. Ranking Algorithms

As shown in Figure8. Tag Rank shows the best MAP and NDCG values and so it could be said that Tag Rank is the best suitable ranking algorithm for this proposed model.

6. Conclusion And Future Work

In this paper, the main aim is to provide new model of Social Network that is based on Multi-Agent Systems concept and the concept of semantic social network. This proposed model mainly consisted of two main agents: indexing agent that carries out enhanced Latent Dirichlet Allocation algorithm (E-LDA), and ranking agent that carries out Tag Rank algorithm.Enhanced LDA (E-LDA) is distinguished from other preceding indexing algorithms and simulation results show an increase precision and recall using E-LDA.E-LDA is enhanced from LDA with (4%), and shows better performance than other semantic indexing algorithms.

Semantic Tag Rank is also distinguished from other ranking agents as it deals with tags that is more relevant to social networks and also more relevant to semantics.

In the future, the term per topic index is suggested to be entered as tags to be processed by ranking agent. This means that we will have larger data to be ranked. So the processing conditions must be taken care of while implementing the system.

A new model of social networks depending on semantics is proposed, with using semantic indexing methods and rank algorithms. In addition, show in test how this idea will be implemented. Then building and implementing the proposed model to a semantic social network can be suggested. Either in an existing social network, or in new semantic social network programmed from the beginning based on the proposed model in this paper.

References

[1] Obar, Jonathan A., Wildman, Steve. Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy. 39 (9): 745–750. doi:10.1016/j.telpol.2015.

[2] Boyd, Dana. Ellison, Nicole. Social Network Sites: Definition, History, and Scholarship, Michigan State University, (2007).

[3] Boyd, Dana; Crawford, Kate. Six Provocations for Big Data. Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. (September 21, 2011). doi:10.2139/ssrn.1926431.

[4] tephen Downes. The Semantic Social Network. February 14, 2004.

[5] Wooldridge, Michael. An Introduction to MultiAgent Systems. John Wiley & Sons. (2002) p. 366. ISBN 0-471-49691-X.

[6] Franchi, Enrico , "A Multi-Agent Implementation of Social Networks", Proceedings of the 11th WOA 2010 Workshop, DagliOggettiAgliAgenti, Rimini, Italy, September 5-7, 2010.

[7] Christopher D. Manning, PrabhakarRaghavan and HinrichSchütze, "Introduction to Information Retrieval". Cambridge University Press. 2008.

[8] Blei, David M.; Andrew Y. Ng; Michael I. Jordan. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3: 993–1022. doi:10.1162/jmlr. 2003.3.4-5.993.2003.

[9] Wang, Y., Lee, J.-S. and Choi, I.-C. "Indexing by Latent Dirichlet Allocation and an Ensemble Model". Journal of the Association for Information Science and Technology, 67: 1736–1750. doi:10.1002/asi.23444. 2016.

[10] The Natural Language Processing Group at Stanford University, https://nlp.stanford.edu/, accessed in October 1, 2017.

[11] Iowa State University, http://home.eng.iastate.edu, accessed in October 10, 2017.

[12] Matlab Topic Modeling Research Toolbox in University of California, Irvine, http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm, accessed in October 12, 2017.

[13] Järvelin, Kalervo and JaanaKekäläinen. "IR evaluation methods for retrieving highly relevant documents." SIGIR Forum 51 (2000): 243-250.

[14] R. Hamamreh and S. Awad, "Tag Ranking Multi-Agent Semantic Social Networks ," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, 2017.

[15] Choi, In-Chan & Lee, Jaesung. "Document Indexing by Latent Dirichlet Allocation". In proceedings of the 2010 international conference on data mining, At Los Angeles, 2010.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[17] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[18] Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[19] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International

AUTHORS

Rushdi A. Hamamreh has PH.D. in Distributed Systems and Networks Security, He graduated at the Saint Petersburg State Technical University in 2002; He is Associate Professor and Head of Computer Engineering at Al-Quds University. His research interests include Networks Security, Routing Protocols, Multiagent Systems and, Cloud and Mobile Computing.

SamehAwad graduated in Computer Engineering in 2008 from Al-Quds University. Since that he has been working in the Department of Information Technology in Birzeit University. In 2018 he has completed a MSc in Electronics and Computer Engineering from Al-Quds University.