FCI-H
Informatics Bulletin

Helwan University

Semantic Tagging-Based Document Retrieval Using Non-Negative Matrix Factorization

Fatma S.Gadelrab[a*], Mohamed H.Haggag[b], Rowayda A.Sadek[c].

*[a]Department of Information Technology. Faculty of Computers and Information Helwan University Helwan, Cairo, Egypt.**

[b]Professor, Department of Computer Science. Faculty of Computers and Information Helwan University, Helwan, Cairo, Egypt.
[c]Assistant Professor, Department of Information Technology. Faculty of Computers and Information, Helwan University, Helwan, Cairo, Egypt.

KEYWORDS

semantic tagging, topic model, semantic document retrieval, non-negative matrix factorization

ABSTRACT

Many document retrieval methods focusing on unstructured text to deliver more meaningful information on the user. Tag-based document retrieval aims to address a challenge to searching relevant text-documents given a set of tags. Tag-based approaches received a wide attention as a possible solution to the big-content related IR, showing a high performance through a combination of its effectiveness and efficiency. This paper use word sense disambiguation with non-negative matrix factorization to generate topic model based semantic.

1. Introduction

With increasing the amount of the data and the emergence of big data, the processing and the analyzing requires the different technology from the earlier. The content Management System like Wikipedia stores and links the huge amount of documents and files. There is a lack of semantic linking and analysis. To reduce the number of references for a selected content, there is a need for semantic matching [18]. Tags (categories or topic) are set of terms serving as a bridge of communication between the user and the documents. The topic modeling has many algorithms like Latent Semantic Analysis (LSA), Probabilistic LSA, and Latent Dirichlet Allocation (LDA) to generate the topic model for tag-based [1].

The word sense disambiguation (WSD) consists of assigning the proper meaning to a word in a certain context. Many pieces of research use the WSD with the topic model instead of word for semantic relation to improve the information extraction [6, 22].

The LDA received much attention in the field of tag-based because of its extensible nature of the model design as a generative process [13]. The Dirichlet distribution has no convincing linguistic motivations and conflicts with two natural assumptions of sparsity: (1) most of the topics have zero probability in a document, and (2) most of the words have zero probability in a topic. Finally, The Bayesian inference complicates the combination of many requirements into a single multi-objective topic model [21],

The contribution of this paper as the following:

The motivation of the proposed solution by adding linguistic knowledge to the representation (such as semantic similarity based information content and word sense disambiguation).

The rest of this paper is organized as follow. In section 2: background, section 3: some recent related work is briefly reviewed, section 4: the proposed solution (Methodology), section 5: result, section 6: conclusion and future work, at last in section 7: references.

* *Corresponding author.* Tel.: +201015508494.
E-mail address: fatma.sayed@fci.helwan.edu.eg

2. Related work

Generally, clustering algorithms can further be classified as hard or soft clustering. Hard clustering computes a hard assignment, each document is a member of exactly one cluster. The assignment of soft clustering algorithms is soft a document's assignment is a distribution over all clusters. In a soft assignment, a document has fractional membership in several clusters. The soft clustering has the following advantages: suppose X is the document can belong to multiple clusters; thus, the user can find multiple themes (tags) for a document X. The measure associated with clusters and documents can be used as a relevance measure to order the document appropriately. Tag-based aims to address a challenge of searching relevant text-documents given a set of tags [12].

2.1. Tagging

Hong, et al. (2017). Proposed a semantic tag recommendation technique exploiting associated words that are semantically similar or related to each other using the inter-wiki links of Wikipedia. The candidate words were then rearranged according to importance by applying a link-based ranking algorithm and then the top-k words were defined as the associated words for the article [6]. The main limitation in this approach, it did not take in consideration the word sense disambiguation. If they used the NMF with relationship graph, may extract inner relation and improved clustering, like Peng, et al. (2017). Proposed a graph regularized a NMF method capable of feature learning and applied it to clustering, meanwhile, the graph of the data is constructed using cleaner features in the feature learning process, which integrates feature learning and manifold learning procedures into a unified NMF model. This method distinguishes features by incorporating a feature-wise sparse approximation error matrix in the formulation [16].

Li, et al. (2016). Developed a Correlated Tag Learning (CTL) model for the semi-structured corpora based on the topic model to enable the construction of the correlation graph among tags via a logistic normal participation process. The outputs of the CTL model was the tags' correlation matrix, and the latent topics for documents [14]. There were two main limitations in this approach. Firstly, it ignored the importance of different tags in a specific document, where some tags were more relevant to a document than others but in another document, the situation could be totally different. Secondly, as described above, the tags were a set of semantic topic distributions, which were learned from the plain text, and so the correlations should be modeled from the semantic level, while only considering the co-occurrences was not enough.

Allahyari, et al. (2016). Proposed a probabilistic topic model that incorporates DBpedia Knowledge into the topic model for tagging Web pages and online documents with topics discovered. They learn the probability distribution of each category over the words using the statistical topic models taking into account the prior knowledge from Wikipedia about the words, and their associated probabilities in various categories [2]. There were two main limitations in this approach. Firstly, did not use linguistic techniques to address annotation of Web resources and did not employ regular expression patterns for semantic tagging. Secondly, take all the words into consideration did not use any dimensionality reduction technique like feature selection or feature extraction. Xu, et al. (2017). A novel knowledge-based topic model, WCM-LDA (Wikipedia-Category-concept-Mention Latent Dirichlet Allocation), was proposed, which not only modeled the relationship between words and topics but also utilizes the concept and category knowledge of entities to model the semantic relation of entities and topics [22].

There were two main limitations in this approach. Firstly, did not use linguistic techniques like WSD of words into topic models to discover more coherent topics? Secondly, used LDA for topic model as mentioned before in [20] LDA has many linguistics problem. NMF preferred to use in text topic model [3].

Li, et al. (2013). Proposed a novel method to model tagged documents by a topic model, called Tag-Weighted Topic Model (TWTM). TWTM was a framework that leverages the tags in each document to infer the topic components for the documents. [13]. There was main limitations in this approach. It did not use WSD and semantic similarity between words to improve tagging.

2.2. Linguistics approach

Many researches applied linguistics like a word sense disambiguation (WSD) with a topic model to build tags based semantics like Izquierdo, et al. (2015). Presents an approach to word sense disambiguation based on the topic Modeling (LDA). This approach consists of two different steps, where first a binary classifier is applied to decide whether the most frequent sense applies or not, and then another classifier deals with the not most frequent sense cases. [7]. There were two main limitations in this approach. It used the LDA for the topic model as mentioned before in [6] the LDA has many linguistics problems. NMF preferred to use in text topic model [7, 3]

Table 1 - Summary of work done by different authors against proposed approach

Reference	WSD	Semantic similarity	NMF
6	X	X	X
16	X	X	X
13	X	X	X
2	X	√	X
22	X	X	X
13	√	√	X
7	X	√	X

3. Proposed model

The goal in this study is to find a topic model based non-negative matrix factorization with lexical semantic using word sense disambiguation (LESK). This section introduces a model that extends non-negative matrix factorization (NMF) for tagging document based topic model. Proposed solution considers tagging as category or set of word sense (terms as key-word). Word sense is feature in document. The proposed model also provides an overview of various NMF extensions and examine their relationships.

$$V=WH$$

NMF equation contains V is document term matrix (TF-IDF) matrix, W is topic document matrix and H is topic term matrix.

3.1. Proposed model pip-line

Fig.1 - Proposed model pip-line

Prepossessing

The aim of this step is getting a first correlation coefficient matrix (H) for NMF. Ordered steps are documents collection, partitioning documents to paragraphs, partitioning documents to sentence, sentence word tokenize, remove stop-words and punctuation, word sense disambiguation by LESK for each remaining word in each sentence to get the bag of synset (sense-word) not bag of word, extracting TF-IDF from bag of Synset (V matrix), remove redundant of bag of Synset, the last is getting a res-correlation coefficient matrix. The following proposed algorithms used in proposed model.

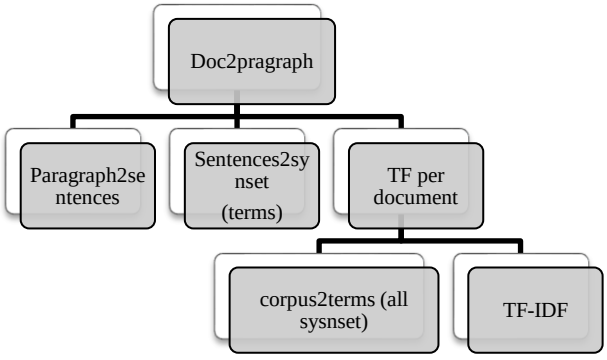

Fig.2 - Prepossessing semantic tagging

1 Doc2Paragraph: splitting documents to paragraphs

2 Paragraph2sentences: splitting paragraph to sentences

3 Sentences2terms: splitting sentence into words (tokenize), removing term noises like stop-words, punctuation, numbers, symbols etc., using word sense disambiguation (LESK) for each remaining word to get the bag of synset (sense-word) not the bag of the word

4 TF per document: compute the terms frequency per the document

5 corpus2terms: remove the frequent terms to produce the collection of terms or the term vector

6 TF-IDF: from step 5 and step 4 compute TF-IDF

Topic model

The aim of this step is to generate a topic model (topic-document and topic-term) based V matrix (TF-IDF) using double singular value decomposition.

4. Experiment and Results

In this section, we corroborate the effectiveness of our model by using 2 dataset

Table 2 - Dataset statistics

Data-set	DOCs size	Category name	Category Docs	Category WN noun	Selected Docs	Selected WN
20-Newsgroup	11314	comp.graphics	584	1023	100	100 n
Reuters (R20)	10788	grain	578	1023	100	100 n

4.1. Experimental setup

Dataset: We use two datasets in the experiments: 20-Newsgroups [1] and Reuters (R20) [2]

20Newsgroups: The 20Newsgroups contains approximately 20,000 newsgroups documents being partitioned (nearly) evenly across 20 different newsgroups, we used the 20newsgroups version downloaded from http://www.ai.mit.edu/~jrennie/20Newsgroups. In our experiments, we used the comp.graphics categories train dataset as sub-set dataset from 20-Newsgroups.

Reuters-21578: The Reuters dataset has been used in many text categorization experiments; the data was collected by the Carnegie group from the Reuters news-wires in 1987. There are now at least five versions of the Reuters datasets widely used in TC community. We

[1] http://qwone.com/~jason/20Newsgroups/

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

choose the Modapte version of the Reuters-21578 collection of new stories downloaded from (5) in our experiments, we used the grain categories (from Reuters (R20).

Their statistics are summarized in table 2. In a preprocessing step, we excluded all the non-content LESK term whose part of speech tags are not noun. We removed stop words hence focusing on relevant content words.

Sub-set dataset: We inspired subset of dataset from [23], subset of dataset train our models and compared models on two data-sets by choosing max 100 features (synset term) in TF-IDF matrix and 100 docs related to these features .

Baselines: We use in our model traditional model (Non-negative Double Singular Value Decomposition (NNDSVD)).

Parameter Settings: For all methods, we learn 100 topics for 100 documents with 100 term feature.

Experiments: We inspire the experiments and evaluation from [23]-[24]. We design two search tasks to test our models and reflect two evaluation (1) Qualitative Evaluation (2) Quantitative Evaluation. The First searches about microphone.n.01 term in 20-Newsgroups dataset and the second searches april.n.01 in Reuter's dataset then show the results by two models (our method-algorithm 4 and traditional NMF)

4.2. Results

We compare our model with the baseline methods qualitatively.

4.2.1. Qualitative Evaluation

Top Topic-terms:

Tables 3 and 4 show some exemplar top 5 topics related to searched term (microphone.n.01 and april.n.01), which is learned by traditional NMF using the two data sets (20-Newsgroups and Reuters). In traditional model, each topic is visualized by the top ten terms. The header cells in each table are names of topics. Table 3 shows search results about microphone.n.0 in 20-nwesgroups dataset. Topic1 (2) in (table 3) has week linguistically relation between terms or no linguistically relation like million.n.01, mistake.n.01. The linguistically relations between terms in one topic or between searched term and resulted topics effects on the result accuracy in semantic search which leads to inaccuracy results in document retrieval. Similarly, in table 4, topic 1(1). No linguistically relation between terms in one topic like aprile.n.01 and peer.n.0.1

Table 3 - Topics-Terms learned from 20-Newsgroups dataset by traditional model

	2	28	99	37	27
0	microphone.n.0	range.n.04	volt.n.01	mode.n.06	phosphorus.n.0
1	range.n.04	cam.n.02	mistake.n.01	mistake.n.01	bible.n.02
2	cam.n.02	microphone.n.0	mortarboard.n.0	dram.n.01	transcript.n.02
3	volt.n.01	volt.n.01	mode.n.06	volt.n.01	space.n.08
4	mistake.n.01	mistake.n.01	fiberglass.n.01	mortarboard.n.0	version.n.06
5	mode.n.06	mode.n.06	million.n.01	fiberglass.n.01	turk.n.01
6	fiberglass.n.01	fiberglass.n.01	turk.n.01	million.n.01	volt.n.01
7	million.n.01	million.n.01	keyboard.n.01	turk.n.01	magnesium.n.0
8	turk.n.01	turk.n.01	gunman.n.02	keyboard.n.01	mode.n.06
9	keyboard.n.01	keyboard.n.01	hexagon.n.01	gunman.n.02	fiberglass.n.01

Table 4 - Topics-Terms learned from Reuters dataset by traditional model

	1	35	31	57	38
0	april.n.01	elevation.n.0	redemption.n	seaway.n.01	worm.n.02
1	hassium.n.01	whitethorn.n.	bushel.n.02	cargo.n.01	peer.n.01
2	pale_yellow.n.01	source.n.07	security.n.04	percentage.n	cargo.n.01
3	redemption.n.02	april.n.01	pale_yellow.	season.n.03	stage_set.n.0
4	nothing.n.01	percentage.n.	april.n.01	april.n.01	pale_yellow.
5	workweek.n.01	loanword.n.0	stage_set.n.0	whitethorn.n	april.n.01
6	worm.n.02	connecticut.n	exporter.n.01	montana.n.0	registration.
7	peer.n.01	billion.n.03	treatment.n.0	worm.n.02	month.n.02
8	government_accounting_off	cooperative.	hectare.n.01	metric_ton.n	exporter.n.0
9	duty.n.03	elevator.n.01	worm.n.02	peer.n.01	cooperative.

Top Topic-doc:

Also qualitative evaluation represents a top 10 document ranked per top 5 selected topic using two dataset for the same experiment in topic-term. Table 5 and 6 show this result .The header cells in each table are names of topics, the index is number of row and the value in each cell is number of document in dataset. These documents are based document topic distribution.

Table 5 - Topics-Docs Learned from 20-Newsgroups dataset by traditional model

	2	28	99	37	27
0	205	67	543	414	479
1	67	414	414	456	342
2	320	479	479	353	62
3	507	507	507	543	194
4	149	149	149	480	241
5	25	25	25	479	455
6	541	541	541	507	543
7	46	46	46	149	507
8	537	537	537	25	149
9	27	27	27	541	25

Table 6 - Topics-Docs Learned from Reuters dataset by traditional model

	1	35	31	57	38
0	235	521	372	527	504
1	66	526	560	225	439
2	299	215	290	374	432
3	271	152	35	20	560
4	524	275	386	504	183
5	57	20	41	532	296
6	268	136	42	176	41
7	30	555	268	524	386
8	12	335	296	261	42
9	372	527	89	16	89

4.3. Summary

For Semantic search, experiments uses the LESK as word sense disambiguation algorithm to determine suitable synset from context and replaces the original word with this synset to compute the TF-IDF. The topic modelling is a resist researching single term in search, and instead move towards exploring term themes. The topic modelling is a complex way that search engines determine that your content is what the search is really looking for, taking into account the other subjects (topics) you discuss in your text. The topic modelling provides us with methods to organize, understand and summarize large collections of textual information. It helps in:

- Discovering hidden topical patterns that are present across the collection based semantics linguistically
- Annotating documents according to these topics
- Using these annotations to organize, search and summarize texts
- Discovering the semantic relation between terms in the same topic (topic-term matrix).

The topic modelling is a kind of ranking system, which decides what context of a particular search term means the most. In order to determine ranking, the search engine measures how far away a key term is from a contextual determinant.

5. Conclusion and future work

Results of the proposed model clear the impact of lexical semantic on accuracy of the document clustering by the proposed non-negative matrix factorization. Also the proposed method to generate the topic model by giving document-term (TF-IDF) and topic-term (correlation coefficient semantic similarity matrix, future, taking into account insatiability problem, dimensionality reduction based semantic relation.

6. References

1.Alghamdi, R., & , K.,2015.A Survey of Topic Modeling in Text Mining.International Journal of Advanced Computer Science and Applications (IJACSA), 6(1).

2. Allahyari, M., & Kochut, K., 2016.Semantic tagging using topic models exploiting Wikipedia category network.InSemantic Computing (ICSC), 2016 IEEE Tenth International Conference on (pp.63-70).IEEE.

3. Belford, M., MacNamee, B., & Greene, D., 2016.Ensemble Topic Modeling via Matrix Factorization.In 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), Dublin, Ireland, 20-21 September 2016 (Vol.1751).CEUR Workshop Proceedings.

4. Belford, M., Mac Namee, B., & Greene, D., 2017. Stability of Topic Modeling via Matrix Factorization. arXiv preprint arXiv:1702.07186.

5. Boyd-Graber, J.L., Blei, D.M., & Zhu, X., 2007.A Topic Model for Word Sense Disambiguation.In EMNLP-CoNLL (pp.1024-1033).

6.Hong, H.K., Kim, G.W., & Lee, D.H., 2017.Semantic tag recommendation based on associated words exploiting the interwiki links of Wikipedia.Journal of Information Science, 0165551517693497.

7. Izquierdo, R., Postma, M., & Vossen, P., 2015.Topic Modeling and Word Sense Disambiguation on the Ancora corpus.Procesamiento del Lenguaje Natural, 55, 15-22.

8. Jindal, R., & Taneja, S.A, 2016.Wordnet Based Semantic Approach for Dimension Reduction in Multi label Text Documents.International Science Press I J C T A, 9(40), pp.267-274

9. Kimura, K., Kudo, M., & Sun, L., 2015.Dimension Reduction Using Nonnegative Matrix Tri-Factorization in Multi-label Classification.In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA) (p.250).The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

10. Kuang, D., Choo, J., & Park, H., 2015.Nonnegative matrix factorization for interactive topic modeling and document clustering.In Clustering Algorithms (pp.215-243).Springer International Publishing.

11. Lee, S., Masoud, M., Balaji, J., Belkasim, S., Sunderraman, R., & Moon, S.J.,2017.A survey of tag-based information retrieval.International Journal of Multimedia Information Retrieval, 6(2), 99-113.

12. Lee, S., Masoud, M., Balaji, J., Belkasim, S., Sunderraman, R., & Moon, S.J.,2017.A survey of tag-based information retrieval.International Journal of Multimedia Information Retrieval, 6(2), 99-113.

13. Li, S., Li, J., & Pan, R.,2013.Tag-Weighted Topic Model for Mining Semi-Structured Documents.In IJCAI (pp.2855-2861).

14.Li, S., Pan, R., Zhang, Y., & Yang, Q., 2016.Correlated Tag Learning in Topic Model.In UAI

15. Pedersen, T., 2010.Information content measures of semantic similarity perform better without sense-tagged text.InHuman Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp.329-332).Association for Computational Linguistics.

16. Peng, C., Kang, Z., Hu, Y., Cheng, J., & Cheng, Q., 2017.Nonnegative matrix factorization with integrated graph and feature learning. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3),

17. Priyadarshini, R., Tamilselvan, L., Khuthbudin, T., Saravanan, S., & Satish, S., 2015.Semantic Retrieval of Relevant Sources for Large Scale Virtual Documents. Procedia Computer Science, 54, 371-379.

18.Ramkumar, A.S., & Poorna, B.2016.Text Document Clustering Using Dimension Reduction Technique.International Journal of Applied Engineering Research, 11(7), 4770-4774.

19.ur Rehman, M.H., Liew, C.S., Abbas, A., Jayaraman, P.P., Wah, T.Y., & Khan, S.U., 2016.Big data reduction methods: a survey.Data Science and Engineering, 1(4), 265-284.

20. Vorontsov, K., & Potapenko, A.,2014.Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization.In International Conference on Analysis of Images, Social Networks and Texts_x000D_ (pp.29-46).Springer International Publishing.

21. Wang, J., Bansal, M., Gimpel, K., Ziebart, B.D., & Clement, T.Y., 2015.A sense-topic model for word sense induction with unsupervised data enrichment. Transactions of the Association for Computational Linguistics, 3, 59-71.

22.Xu, K., Qi, G., Huang, J., & Wu, T., 2017.Incorporating Wikipedia concepts and categories as prior knowledge into topic models.Intell.Data Anal., 21(2), 443-461.

23.Guo, W., & Diab, M. (2011, July). Semantic topic models: Combining word distributional statistics and dictionary definitions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 552-561). Association for Computational Linguistics.

24.Xie, P., Yang, D., & Xing, E. (2015). Incorporating word correlation knowledge into topic modeling. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 725-734).