SIMT: A Semantic Interest Modeling Toolkit

Mohamed Amine Chatti
University of Duisburg-Essen
Germany
mohamed.chatti@uni-due.de

Fangzheng Ji
University of Duisburg-Essen
Germany
fangzheng.ji@stud.uni-due.de

Mouadh Guesmi
University of Duisburg-Essen
Germany
mouadh.guesmi@stud.uni-due.de

Arham Muslim
National University of Sciences and Technology
Pakistan
arham.muslim@seecs.edu.pk

Ravi Kumar Singh
RWTH Aachen University
Germany
ravi.singh@rwth-aachen.de

Shoeb Joarder
University of Duisburg-Essen
Germany
shoeb.joarder@uni-due.de

ABSTRACT

In this paper, we focus on semantic interest modeling and present SIMT as a toolkit that harnesses the semantic information to effectively generate user interest models and compute their similarities. SIMT follows a mixed-method approach that combines unsupervised keyword extraction algorithms, knowledge bases, and word embedding techniques to address the semantic issues in the interest modeling task.

KEYWORDS

Interest modeling, Keyword extraction, Semantic similarity, Word embedding, Interest embedding, Interest model embedding

ACM Reference Format:
Mohamed Amine Chatti, Fangzheng Ji, Mouadh Guesmi, Arham Muslim, Ravi Kumar Singh, and Shoeb Joarder. 2021. SIMT: A Semantic Interest Modeling Toolkit. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21 Adjunct), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3450614.3461676

1 INTRODUCTION

Semantics-aware content representation is crucial for the success of adaptive and personalized systems, such as recommender systems [15]. In this paper, we apply semantics in the user interest modeling task and present a Semantic Interest Modeling Toolkit (SIMT) for the effective generation and similarity computation of interest models, based on semantic information. Inspired by the fact that an interest is often a well-defined concept (article) in Wikipedia, we introduce a method of interest modeling that mixes unsupervised keyword extraction algorithms and Wikipedia as a knowledge base to generate semantically-enriched interest models. Moreover, we propose new similarity measures that leverage Wikipedia and word embedding techniques to address the semantic issues in the computation of similarity between interest models.

UMAP '21 Adjunct, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8367-7/21/06.
https://doi.org/10.1145/3450614.3461676

In the demo, we will show how SIMT has been leveraged in the transparent recommendation and interest modeling application (RIMA)[1] to infer interest models of researchers based on their publications extracted from Semantic Scholar[2] and use the inferred interest models to provide personalized recommendations of tweets.

2 RELATED WORK

2.1 Interest Model Generation

Different text mining techniques have been used to generate user interest models. These include text classification [13, 25, 32], named-entity recognition [17, 26], and keyphrase extraction, which is widely used to generate interest models from text-based and social media data [2, 11]. In this work we focus on keyphrase extraction methods. The existing approaches can be categorized based on the type of the method (e.g., supervised, unsupervised) used to generate interest keyphrases/keywords [29]. In supervised approaches, classification algorithms are commonly used to allocate users to predefined interest classes based on their data [27]. Even though supervised approaches are relatively simple and easy to apply, they are domain-dependent and limited in identifying only predefined interests which were used to train the prediction model.

Unsupervised approaches, on the other hand, are not dependent on any predefined prediction model. Thus, they can generate a more diverse set of user interests. Unsupervised keyword extraction approaches can be further divided into statistical-based (e.g., LDA [16, 26], TF-IDF [31], Rake [28], YAKE! [8]) and graph-based approaches (e.g., TextRank [19], PageRank [22], SingleRank [34], TopicalPageRank [14], TopicRank [7], MultipartitieRank [6], PositionRank [9]).

These supervised and unsupervised approaches, however, do not consider the semantic issues during the keyword extraction phase. In order to address this issue, some research works incorporated knowledge bases such as DBPedia [24], Freebase [36], Linked Open Data (LOD) cloud [1], YAGO [30], and Wikipedia [4, 12, 17, 18, 21, 33] to infer user interest models. Our approach for interest model generation goes beyond existing works by mixing unsupervised keyword extraction algorithms and Wikipedia as a knowledge base to generate semantically-enriched domain-independent user interest models.

[1] https://rima.sc.inko.cloud/
[2] https://www.semanticscholar.org/

2.2 Interest Model Similarity

To the best of our knowledge, there are no research works that focus on finding semantic similarity between interest models. However, semantic similarity has been introduced and applied by many researchers in the NLP community through (a) knowledge-based algorithms and (b) corpus-based algorithms. Knowledge-based similarity algorithms focus on identifying the degree of similarity between words using information derived from semantic networks [10]. For example, the Wiki-Link Measure (WLM) uses the Wikipedia's hyperlink structure rather than its category or text content to compute the similarity between words/concepts [30, 35].

Corpus-based algorithms compute similarity between words by using large corpora. Thereby, semantic representations of the words are needed. Word embeddings represent a corpus-based semantic representation method which has been widely used in various NLP applications [3, 20]. Examples of word embedding techniques include Word2Vec [20], GloVe [23], and FastText [5].

In our work, we apply word embedding methods to represent the generated interest models. Also, since we are generating the interest model using Wikipedia, we propose another semantic method based on the Wiki-Link Measure (WLM). Based on these two methods, we can use the cosine measure to calculate the similarity between interest models.

3 SEMANTIC INTEREST MODELING TOOLKIT

The Semantic Interest Modeling Toolkit (SIMT)[3] aims at using the semantic information to effectively generate interest models and compute their similarities. SIMT consists of two main components, namely 'Interest Model Generation' and 'Semantic Interest Model Similarity'. These components are developed as RESTful APIs allowing them to be easily used by any personalized system that requires semantic user interest modeling.

3.1 Interest Model Generation

This component is responsible for generating a user's interest model. It contains two sub-components: 'Keyword Extractor' and 'Semantic Enrichment'. The 'Keyword Extractor' sub-component is responsible for extracting candidate interest keywords from the user-generated textual content (e.g. social media posts, research publications) using various unsupervised keyword extraction algorithms including TextRank, SingleRank, TopicRank, TopicalPageRank, PositionRank, MultipartitieRak, Rake, and YAKE!. These algorithms are employed from three different open-source python libraries for keyword extraction, namely pke[4], yake[5], and rake-nltk[6]. The 'Semantic Enrichment' sub-component leverages semantic information from Wikipedia to generate the user's interest model based on the candidate interest keywords generated from the 'Keyword Extractor' sub-component. Leveraging Wikipedia to infer interest models has the potential to address different semantic-related issues: (a) synonym interests can be merged (b) acronym interests can be reduced, and (c) noise coming from non-relevant keywords can be filtered out. As a result, Wikipedia-based interest models would be more representative and more accurate than keyword-based ones [21]. Moreover, enriching an interest model with Wikipedia categories might lead to unexpected but still relevant interests, which can be important to achieve serendipity in recommender systems. The three main steps of the interest model generation process are described below:

Keyword-based Interest Model

(1) Use different preprocessing techniques, such as data cleaning, transformation, reduction, and integration to prepare the data (textual content) for keyword extraction
(2) Apply an unsupervised keyword extraction algorithm on the preprocessed data to get candidate interest keywords
(3) Merge extracted keywords and assign them a weight based on their occurrences. Afterwards, normalize the weights to generate a keyword-based interest model of the user

Wikipedia-based Interest Model

(4) Leverage Wikipedia as a lexical knowledge base to map the candidate interest keywords to their linked Wikipedia articles. That is, if there exists a Wikipedia article for the keyword, then the article's title is added to the Wikipedia-based interest model, otherwise the keyword is discarded
(5) Identify redirected keywords, which do not have their own Wikipedia articles but they are linked to some other articles. Then, merge the weights of all the keywords redirecting to the same Wikipedia article in order to improve the Wikipedia-based interest model by reducing redundancy
(6) Normalize the weights of the Wikipedia-based interest model

Wikipedia Category-based Interest Model

(7) Select the top three interests in the Wikipedia-based interest model
(8) Generate a Wikipedia category-based interest model of the user using the categories of the selected top three interests from the Wikipedia-based interest model

Following the steps of the interest model generation process, keyword-based and Wikipedia-based interest models for one author were generated based on his research publications in the last five years. By comparing the keyword-based interest model (Figure 1) and the Wikipedia-based interest model (Figure 2), it can be observed that: (a) synonym interests are merged: *learning* and *learning process* are merged to *Learning*, *peer assessment* and *peer assessment process* are combined under *Peer assessment*, (b) acronym interests are reduced: *moocs* is replaced with *Massive open online course*, and (c) less interesting keywords are removed: keywords having no linked Wikipedia articles are discarded, e.g. *educational data sets*.

3.2 Semantic Interest Model Similarity

This component is responsible for calculating semantic similarity scores between two interest models. The first step in calculating the similarity is to generate a vector representation of both models. Afterwards, the similarity is calculated by applying cosine similarity to the two interest model vectors. SIMT computes the semantic

[3] An open-source implementation of SIMT is available at https://github.com/udesoco/RIMA
[4] https://github.com/boudinfl/pke
[5] https://pypi.org/project/yake/
[6] https://pypi.org/project/rake-nltk/

Figure 1: Top 15 interests in the keyword-based interest model of an author generated using SingleRank

Figure 2: Top 15 interests in the Wikipedia-based interest model of an author

similarity of interest models using two different approaches, namely *Wikipedia-based Measure* and *Word Embedding-based Measure*.

Wikipedia-based Measure. This approach uses the Bag-of-Words (BoW) technique together with the Wiki-Link Measure (WLM) described in [35] to first generate vector representations of two interest models and then compute the cosine similarity between the two vectors, as shown in Figure 3. The steps to compute the semantic similarity between two interest models using the *Wikipedia-based Measure* are described below:

(1) Generate an interest vector space containing all the interests from both interest models
(2) Represent each interest model with an interest vector and assign a score of '1' to all the keywords in the vector space which are also present in the interest model. For each keyword w in the vector space which is not present in the interest model, use the WLM to calculate the semantic relatedness sr between w and all the other keywords in the interest model having a score of '1'. Then, take an average of all the calculated semantic relatedness results and assign it as the score of w in the interest vector
(3) Calculate the cosine similarity between the two interest vectors

Word Embedding-based Measure. This approach uses word embedding techniques to generate *interest embeddings* (i.e. word embedding representations of interest) and *interest model embeddings* (i.e. word embedding representations of interest models). We use GloVe [23], as it is trained on Wikipedia data. The steps to compute the semantic similarity between two interest models using the *Word Embedding-based Measure* are shown in Figure 4 and described below:

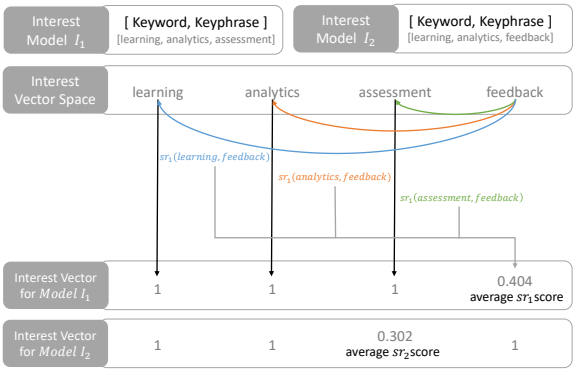

Figure 3: Wikipedia-based Measure

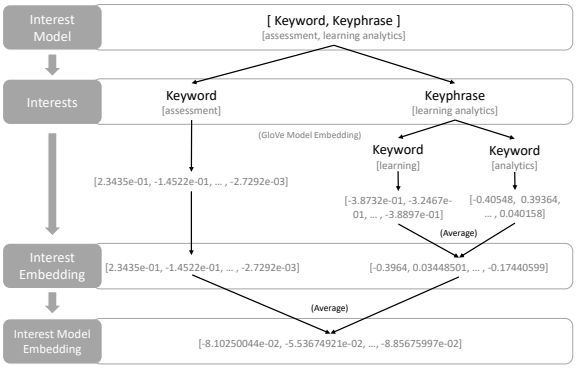

Figure 4: Word Embedding-based Measure

(1) Load a GloVe pre-trained word embedding model
(2) Collect all interests from an interest model along with their respective weights. If an interest is a keyphrase, split it into keywords
(3) Use GloVe to create an interest embedding for each interest. If an interest is a keyphrase, take the average of the interest embeddings of each keyword in the keyphrase
(4) Compute the interest model embedding as the weighted average value of all the interest embeddings
(5) Calculate the cosine similarity between the two interest model embeddings

4 CONCLUSION AND FUTURE WORK

In this demo paper, we presented a Semantic Interest Modeling Toolkit (SIMT) that harnesses the semantic information based on Wikipedia and word embedding techniques for the effective generation and similarity computation of interest models. As future work, we will conduct experiments to compare our approach with traditional keyphrase extraction and similarity computation approaches in the interest modeling task. Further, we plan to consider and evaluate other embedding techniques such as BERT and ELMo to compute the semantic similarity between user interest models.

REFERENCES

[1] F. Abel, C. Hauff, G. Houben, and K. Tao. 2012. Leveraging User Modeling on the Social Web with Linked Data. In *ICWE*.

[2] Slobodan Beliga, Ana Meštrović, and Sanda Martinčić-Ipšić. 2015. An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences* 39, 1 (2015), 1–20.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[4] Christoph Besel, Jörg Schlötterer, and Michael Granitzer. 2016. On the quality of semantic interest profiles for onine social network consumers. *ACM Sigapp Applied Computing Review* 16 (2016), 5–14.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[6] Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. *arXiv preprint arXiv:1803.08721* (2018).

[7] Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction.

[8] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! collection-independent automatic keyword extractor. In *European Conference on Information Retrieval*. Springer, 806–810.

[9] Corina Florescu and Cornelia Caragea. 2017. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1105–1115.

[10] Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68, 13 (2013), 13–18.

[11] Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1262–1273.

[12] Ludovic Jean-Louis, Michel Gagnon, and Eric Charton. 2013. A knowledge-base oriented approach for automatic keyword extraction. *Computación y Sistemas* 17, 2 (2013), 187–196.

[13] Kangmiao Liu, Wei Chen, Jiajun Bu, Chun Chen, and Lijun Zhang. 2007. User modeling for recommendation in blogspace. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops*. IEEE, 79–82.

[14] Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 366–376.

[15] Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2019. *Semantics in Adaptive and Personalised Systems*. Springer.

[16] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 889–892.

[17] Matthew Michelson and Sofus A Macskassy. 2010. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. ACM, 73–80.

[18] Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 233–242.

[19] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[21] F. Narducci, C. Musto, G. Semeraro, P. Lops, and M. Degemmis. 2013. Leveraging Encyclopedic Knowledge for Transparent and Serendipitous User Profiles. In *UMAP*.

[22] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[24] Guangyuan Piao and John G Breslin. 2016. User modeling on Twitter with WordNet Synsets and DBpedia concepts for personalized recommendations. In *proceedings of the 25th ACM international on conference on information and knowledge management*. 2057–2060.

[25] Bayu Yudha Pratama and Riyanarto Sarno. 2015. Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 170–174.

[26] Xiao Pu, Mohamed Amine Chatti, Ulrik Schroeder, et al. 2016. Wiki-lda: A mixed-method approach for effective interest mining on twitter data. In *Proceedings Of The 8Th International Conference On Computer Supported Education, Vol 1 (Csedu)*. Scitepress, 426–433.

[27] Mandyam Annasamy Raghuram, K Akshay, and K Chandrasekaran. 2016. Efficient user profiling in twitter social network using traditional classifiers. In *Intelligent systems technologies and applications*. Springer, 399–411.

[28] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory* 1 (2010), 1–20.

[29] Nacéra Bennacer Seghouani, Coriane Nana Jipmo, and Gianluca Quercini. 2019. Determining the interests of social media users: two approaches. *Information Retrieval Journal* 22, 1-2 (2019), 129–158.

[30] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. 2013. Linking named entities in tweets with knowledge base via user interest modeling. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 68–76.

[31] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28, 1 (1972), 11–21.

[32] Mia Stern, Joseph Beck, and Beverly Park Woolf. 1999. Naive Bayes classifiers for user modeling. *Center for Knowledge Communication, Computer Science Department, University of Massachusetts* (1999).

[33] Giorgia Di Tommaso, Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2018. Wiki-MID: A Very Large Multi-domain Interests Dataset of Twitter Users with Mappings to Wikipedia. In *International Semantic Web Conference*.

[34] Xiaojun Wan and Jianguo Xiao. 2008. Single Document Keyphrase Extraction Using Neighborhood Knowledge.. In *AAAI*, Vol. 8. 855–860.

[35] I.H. Witten and D. Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*. 25–30.

[36] Xiao Yu, Hao Ma, Bo-June Hsu, and Jiawei Han. 2014. On building entity recommender systems using user click log and freebase knowledge. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 263–272.