Academic Collaboration Recommendation Based on Sparse Distributed Representation

Zhong Li\textsuperscript{1,a}, Hong-Qi HAN\textsuperscript{1,*}, Guang-Yin WU\textsuperscript{1,2,b}, Xiao-Rui ZHAII\textsuperscript{1,c}

\textsuperscript{1}Research Center for Information Science Theory and Methodology, Institute of Scientific and Technical Information of China(ISTIC), 15 Fuxing Road, Haidian District, Beijing 100038, P.R. China
\textsuperscript{2}Beijing Infosoft Software Co., Ltd. 15 Fuxing Road, Haidian District, Beijing 100038, P.R. China
\textsuperscript{a}lizhong2016@istic.ac.cn, \textsuperscript{b}gywu@wanfangdata.com.cn, \textsuperscript{c}zhaixr2017@istic.ac.cn

Keywords: Collaboration Recommendation; Sparse Distributed Representation; Social network.

Abstract. The researchers with similar research contents are helpful for understanding the relevant field and the promotion of further exchanges and collaborations. Find proper researchers is a complex task, we proposed an innovative approach of collaboration recommendation based on sparse distributed representation (SDR). According to the sparse distributed representation theory, the text contents of authors' thesis are characterized to vectors, then recommend collaborations based on the similarity of vectors. We also validate our approach using the data set of NIPS, the experimental results show that the method we proposed is effectively used for collaborative recommendation.

Introduction

With the development of Internet, people can obtain various information easily, but it is hard to get the information they exactly need. Researchers also need information about researchers in their research field from the vast literature, and the professionalism and complexity of scientific research make collaboration more important\cite{1}. Therefore, there is a need for a method to help researchers quickly find target authors.

Academic collaboration recommendation is a technology to solve such problems. It uses existing scientific and technological information resources such as papers and patents to recommend related researchers. This method can help quickly discover and understand the relevant researchers and research content in the field, promote further exchanges and cooperation, bring better knowledge and resource sharing, improve the quality of scientific research, accelerate the research process, and achieve greater scientific research results.

Hara et al. \cite{2} divided academic collaboration or collaborators into two categories. One refers to the close cooperation between each other in the process of verifying each other's hypotheses, mutual respect, trust, and common development, usually belonging to similar research fields; another emphasizes the complementarity of knowledge and skills between each other, interdisciplinary collaboration mostly falls into this category.

In this paper, we study the first type of collaboration. At present, there are two main types of recommendation methods based on existing author collaboration and authors' research contents. The former method is generally suitable for connected networks, the predictive effect on disconnected networks is relatively poor, and the recommendation is based on the actual social relationship between researchers. The latter method based on research content usually uses TF-IDF \cite{3} or topic model \cite{4} to recommend, the performance is often affected by keywords, feature selection, preferences. Most of them need to train different models for different corpora, and usually have high complexity and low efficiency, which is not suitable for processing large data.

Given the problems existing in the existing academic collaboration recommendation methods, we propose a method based on Sparse Distributed Representation (SDR) \cite{5}. It gets the SDR vector representing the author's research interest, then calculates the similarity of the author's SDR for
collaborative recommendation. This method can train a large-scale corpus in advance, and can represent a paper text quickly and conveniently as a fixed-bit binary vector, so that an author's SDR representation can be obtained quickly and conveniently, so it can be easily switched. It is highly efficient and adaptable to large data environments in different areas.

**Content-based Collaboration Recommendation**

The content-based approach is based on the expertise of researchers and uses data or text mining techniques to make recommendations. The basic idea is to use the textual information representation approaches to express the author's research content in various ways, so as to obtain the similarity between the authors and to recommend according to the similarity.

In order to extract keywords containing certain semantic information, Lee [6] used the Mesh terms to extract the keywords, then he combined TF-IDF with author contribution weights to perform author similarity calculation and recommendation. Guanyao Du et al. [7] proposed a scientific user interest detection method based on the LDA (Latent Dirichlet Allocation) model, which combines cross-collaboration-domain features, and then he presented the collaborators and the article recommendation algorithms based on the method given. Fang H and C Zhai [8] established a probabilistic model for expert recommendation and proved its effectiveness through experimental data sets; Balog et al. [9] provided two different strategies to seek experts in corporate corpora; Deng et al. [10] proposed three content-based expert recommendation methods with reference to Balog’s document-centric approach. The above methods use different ways to embody the semantic relationships, but generally have higher complexity and poorer adaptability.

**Sparse Distributed Representation**

Sparse Distributed Representation (SDR) is an information coding method that mimics the human cerebral cortex proposed by Jeff Hawkins [11], which provides information on the entire cortex and Various functions are encoded [12]. When cortex receives an external stimulus at a certain time, only a small number of neurons are active, while the rest are relatively inactive, so this activity is sparse (due to the presence of inhibitory neurons); and information is not just encoded in a certain neuron, but across a group of active neurons, so this representation is distributed.

In sparse distributed representation, each word received by the sensory system corresponds to a statement, consisting of one or more sentence. Eventually, every word will get linked to more and more new context, strengthening its conceptual grounding.

The specific SDR process is as follows:

1) Definition of a reference text corpus of documents that represents the semantic universe the system is supposed to work in.
2) Every document from the corpus is cut to text snippets with each snippet representing a single text.
3) The reference snippets are distributed over a 2D matrix in a way that snippets with similar topics (share many common words) are placed closer to each other on the map, and snippets with different topics are places more distantly.
4) For each word in the corpus, all the contexts the word occurs in are set to “1” in the corresponding bit-position of a 2D mapped vector.

This produces a large sparse binary vector called word SDR. For a text, we can get text-SDR by combining the SDRs of words in the text, keeping the “1” bit which occurred frequently. Obviously, more frequently the bit value “1” appear, more important the semantics it contain.

Let an n-dimensional SDR vector consisting of binary numbers 0, 1 be represented as $x = [b_1, b_2, ..., b_n]$. which is large and sparse. Although only a small fraction of the x component is “1”, it is still highly representative because its high dimensionality. Use $w_x$ to represent the number of “1” in x, given n and w, the number of SDR codes that are different from each other is...
If \( n = 2048 \), \( w = 40 \), there are \( 2.37 \times 10^{64} \) different SDR codes, and there are only about \( 10^{80} \) atoms visible in the universe. The SDR vector has a strong representation capability if a sufficiently large \( n \) and a suitable \( w \) are selected.

The more “1” in two SDR vectors share the same position, the more semantically they are. Therefore, in addition to Jaccard similarity, cosine similarity, Euclidean distance, etc., it is also possible to directly calculate similarity with overlaps of two SDRs.

Let \( x \) and \( y \) be two SDR vectors, and the overlap can be calculated by the dot product of the vectors:

\[
\text{OVERLAP}(x, y) = x \cdot y
\]

Two SDR vectors are matched when their overlap exceeds a certain threshold \( \theta \). When the threshold is reduced, it means more noise, but the SDR mode has excellent noise immunity. Literature [5] analyzed its anti-noise ability. For example, \( n = 1024 \), \( w = 4 \). If the threshold = 2, corresponding to 50% of the noise, the probability of mismatching is \( 1/14587 \), there is still a high probability that it occurs, but when \( n \), \( w \) increases to 20 and 10 respectively, the corresponding error matching possibility is reduced to \( 1/10^{13} \). Therefore, with a fixed \( n \), SDR can achieve perfect robustness to 50% noise by a small increase in \( \theta \) and \( w \).

**Academic Collaboration Recommendation**

The method is divided into 3 steps: generating authors’ SDR, calculating similarities, and recommending.

**Author’s SDR**

In the academic cooperation recommendation, author's research content is reflected in the author's papers. Therefore, we generate the SDR of an author based on the author’s papers published. The processes are as follows:

1) Generate text SDRs for each article of the author. For convenience of storage, we only store the positions of “1” which is called the position vector;

2) For an author, set \( n \) articles published, generate the position vectors for each article as \( p_{01}, p_{02}, \ldots, p_{0n} \);

3) Put the \( n \) position vectors together in one set, keep \( k \) position numbers which occurred most frequently and remove the rest to keep the sparsity below the threshold we set (i.e. \( k/n \), \( k \): the number of “1” in position vector, \( n \): the length of SDR).

Use the vector obtained above as the author SDR which always stored as a position vector. This method not only retains the semantic information of the author's research content, but also ensures strong anti-noise ability and reduces the probability of mismatching.

**Similarity and Recommendation**

We obtain the similarity between the authors by calculating their SDRs’ similarity. All the similarity metrics we introduced above can be used in this calculation. Then the top \( N \) authors with high similarity are regarded as the recommended results by sorting the similarities from high to low.

Given a paper dataset, the specific processes of calculating the similarity and recommendation are as follows:

1) According to the publication time of the papers, the data set is divided into two parts: the training set and the test set;

2) For each author in the training set, get his/her SDR vector from his/her papers;
3) Build a collaborative network based on the partnership between authors in the training set to identify each author's collaborators and current non-collaborators in the training set, remove the nodes (authors) only appeared in training set but not in test set;

4) For an author A, calculate similarities with all his/her non-collaborators, and sort the results high to low

5) The top N authors will be recommended for A.

**Experiment and Results**

We use NIPS\(^1\) paper 1\(^{st}\) to 10\(^{th}\) as the experiment dataset, the first 8 years were used for training and the last 2 years were used for testing. Among them, there are 984 papers, 1281 authors, 1866 collaborations, and 2.29 authors per paper in average in the training set, while the test set has 304 papers, 519 authors, and 529 collaborations, the average number of authors per paper is 2.23. The 198 authors were all in the training set and in the testing set, and 79 authors had new collaborations, resulting in 52 times.

After that, we use the Retina API\(^2\) to get the SDRs of full texts of papers, then we perform the experiment by using the collaboration recommendation approach we proposed.

| Author | Recommendation | Similarity (overlap) | Co-authors in testset |
|--------|----------------|----------------------|-----------------------|
| Hasler P | Koch C | 733 | Koch C |
| | Murray A | 727 | |
| | Sejnowski T | 722 | |
| | Andreou A | 715 | |
| | Platt J | 711 | |
| Lewicki M. | Baldi P | 1015 | Sejnowski T |
| | Sejnowski T | 1013 | |
| | Guyon I | 1001 | |
| | Pearson J | 997 | |
| | Abu-Mostafa Y | 995 | |

Partial result shows in Table 1. We recommend 5 authors for each author, for Hasler P, Koch C is in the top 5 authors we recommended, who has the highest similarity with Hasler P; other authors recommended also have high similarities with Hasler P, who didn’t co-authored with Hasler P in testing years, but the results also are useful for understanding domain research. Because there are only 52 new collaborations in testing set, there is only 1 new collaboration for most authors.

| Author | Recommendation | Similarity (overlap) | Co-authors in testset |
|--------|----------------|----------------------|-----------------------|
| Hasler P | Koch C | 733 | Koch C |
| | Murray A | 727 | |
| | Sejnowski T | 722 | |
| | Andreou A | 715 | |
| | Platt J | 711 | |
| Lewicki M. | Baldi P | 1015 | Sejnowski T |
| | Sejnowski T | 1013 | |
| | Guyon I | 1001 | |
| | Pearson J | 997 | |
| | Abu-Mostafa Y | 995 | |

We also compare the hits numbers with LDA and TF-IDF, Table 2 shows the results. We choose N=2,5,12 as the recommended number of each author, SDR approach has the best performance for all Ns.

| | N=2 | N=5 | N=12 |
|----------------|------|------|------|
| SDR hits | 6 | 6 | 14 |
| LDA hits | 2 | 7 | 14 |
| TF-IDF hits | 4 | 5 | 10 |

---

1. https://nips.cc/
2. http://www.cortical.io
Conclusion

This paper uses the method of sparse distributed representation (SDR) to academic collaboration recommendation. The sparse distributed representation method considers the semantic relationships among words in the text. Each bit of the generated SDR vector has a specific semantics, and SDR has high-dimensional and sparse features, good representation and anti-noise ability. Through experiments, it is an effective method for academic collaboration recommendation.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China: Study on the Disconnected Problem of Scientific Collaboration Network.

References

[1] Wagner C. The Royal Society Knowledge, Networks and Nations: Global Scientific Collaboration in the 21st Century. https://royalsociety.org/topics-policy/projects/knowledge-networks-nations/report/

[2] Hara, Noriko, et al. "An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration." Journal of the Association for Information Science & Technology 54.10(2010): 952-965.

[3] Kanfer A., Sweet J., Schlosser A. Humanizing the Net: Social Navigation with a “know-who” Email Agent. In Proceedings of the Third Conference on Human Factors and the Web. New York: ACM, (1997).

[4] Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation [J]. J Machine Learning Research Archive, (2003), 3: 993–1022.

[5] Webber F. D. S. Semantic Folding Theory And its Application in Semantic Fingerprinting [J]. Computer Science, (2015).

[6] Lee D. H., Brusilovsky P., Schleyer T. Recommending collaborators using social features and MeSH terms [J]. Proceedings of the American Society for Information Science & Technology, (2011), 48(1): 1–10.

[7] Guanyao Du, Yuchun Liu, Jianjun Yu. Scientific Users’ Interest Detection and Collaborators Recommendation. New York: IEEE(2018).

[8] Fang H., Zhai C. X. Probabilistic Models for Expert Finding. Springer Berlin Heidelberg, (2007).

[9] Balog K., Azzopardi L., Rijke M. D. Formal models for expert finding in enterprise corpora. J. (2006): 43–50.

[10] Deng H., King I., Lyu M. R. Formal Models for Expert Finding on DBLP Bibliography Data. Eighth IEEE International Conference on Data Mining. New York: IEEE, (2008): 163–172.

[11] Hawkins J., Blakeslee S. On Intelligence [M]. St. Martin's Griffin, (2004).

[12] Ahmad S., Hawkins J. Properties of Sparse Distributed Representations and their Application to Hierarchical Temporal Memory. J. Eprint Arxiv, (2015).