Research Hypothesis Generation Using Link Prediction in a Bipartite Graph

Jung-Hun Kim
Department of Industrial and System Engineering
KAIST
Daejeon, South Korea
junghunkim@kaist.ac.kr

Aviv Segev
Department of Computer Science
University of South Alabama
segev@southalabama.edu

***Abstract*—The large volume of scientific publications is likely to have hidden knowledge that can be used for suggesting new research topics. We propose an automatic method that is helpful for generating research hypotheses in the field of physics using the massive number of physics journal publications. We convert the text data of titles and abstract sections in publications to a bipartite graph, extracting words of physical matter composed of chemical elements and extracting related keywords in the articles. The proposed method predicts the formation of new links between matter and keyword nodes based on collaborative filtering and matter popularity. The formation of links represents research hypotheses, as it suggests the new possible relationships between physical matter and keywords for physical properties or phenomena. The suggested method has better performance than existing methods for link prediction in the entire bipartite graph and the subgraph that contains only a specific keyword, such as 'antiferromagnetism' or 'superconductivity.' Our suggested concept for generating research hypotheses can be easily extended to various other research topics or industrial topics using patent literature.**

***Keywords*-Hypothesis Generation; Text Mining; Link Prediction; Graph; Recommender Systems;**

I. Introduction

Automatic methods enabled by algorithms and high-performance computing can generate aggregate level insights that would not otherwise be uncovered by looking at data silos independently. We suggest a method for generating research hypotheses by extracting knowledge from the massive amounts of published literature growing at an exponential rate [1]. Wallas [2] has suggested that generating new ideas is based on 'Incubation,' which represents the subconscious without deliberate focus, and the 'Illumination' phase, which represents a sudden flash of light. Because the process of generating ideas is vague, automated generation of hypotheses is a valuable tool that assists researchers in generating ideas. There are several previous works that generated hypotheses automatically in biology using massive data from literature and experiments [3], [4]. We expand the field to physics, especially condensed matter physics using bipartite graph, collaborative filtering, and matter popularity.

Condensed matter physics is one of the hottest research fields to understand the behaviors or properties of matter (e.g., Graphene, Silicon, FeSe) in various conditions. Some special behaviors or properties have a name like 'superconductivity,' 'Bose-Einstein condensate (BCS)' or 'antiferromagnetism' and they are normally important keywords in the abstract section of the papers.

The proposed model suggests the new research ideas in condensed matter physics based on the relations between keywords and matter in the papers. Publications from 2004 to 2016 in the Physical Review B (PRB) journal and the Physical Review Letter (PRL) journal, which are one of the representative journals for condensed matter physics, were used for the model. We extract words of matter and keywords and construct the bipartite graph using two types of nodes, matter and keywords, and edges or links which are formed when the matter and keywords appear in the title or abstract of the same article.

Predicting the formation of new edges between nodes of matter and keywords represents that the two entities will co-occur in future literature in this research area. The new edges indicate the new relationships between matter and keywords and they contain new ideas which have not been considered previously. For predicting the formation of links, the proposed method uses collaborative filtering (CF) algorithms with popularity of matter from appearance frequency in the publications.

Among the keyword nodes in the bipartite graph, we focus on 'antiferromagnetism' and 'superconductivity'. Antiferromagnetism is one of the magnetic properties in matter applied to reading elements of hard-dist heads and superconductivity is one of the hottest research topics in condensed matter physics with various applications. The prediction of links between matter and those keyword nodes represents that we can predict matter that will be revealed to have a new relationship with those specific keywords.

In this paper, we suggest a method for generating research hypotheses in condensed matter physics and the method shows improved performance for predicting links in a bipartite graph in comparison with benchmark recommendation algorithms.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

II. Related Work

A. Link Prediction in a Graph

Link prediction in a graph is an active research area in computer science. Normally the type of graph is a unipartite graph such as a social network, web pages, and citation network. Liben-Nowelly [5] suggested the idea for link prediction in the co-authorship network for predicting future interactions between researchers using measurements of network topologies. The recommendation problem can be seen as a link prediction in a bipartite graph. In the case of link prediction for the bipartite graph, there is previous work using CF algorithms, graph measures, and graph kernel-based machine learning [6]. However, the algorithms did not consider the characteristics of the domain.

B. Research Hypothesis Generation

There have been efforts in biology to make systems that generate research hypotheses by using text mining in the scientific literature of Medline abstracts or using algorithms for analyzing DNA data [3], [4]. Spangler [3] constructed a system that can find the new protein kinases with target function using graph-based diffusion of information. In genetics, King et al. [4] applied a system to the determination of the gene function. However, previous works are normally limited to the field of biology and the methods are limited to a very specific purpose and hard to be generalized.

Our work extends the application of link prediction in a bipartite graph to generating research hypotheses in physics and suggests an improved method for link prediction considering the characteristics of the domain.

III. Methodology

A. Construction of the Bipartite Graph

For constructing matter nodes, we extract words of matter from the titles of publications but not from abstracts because we only consider the significant physical matter in each paper. The following describes the text patterns used to extract words of matter in titles:

- There is matter which is composed of the list of chemical elements and numbers (e.g., TiSe2, Si(111), FeSe).
- There is matter which includes character 'x' or 'y' (e.g., BaFe2(As1-xPx)(2), FeTe1-xSex, InxGa1-xAs1-yNy).
- There is matter which includes some words 'delta,' 'beta,' 'alpha,' 'doped' and 'based' (e.g., BiS2-based, alpha-FeTe, beta-CaCr2O4).
- There is matter which includes notation '/' (e.g., Co/Cu, InAs/GaAs, Si/Ge).
- There are special materials which have a name themselves (e.g., graphene, silicone, diamond).

Lastly, we remove the trivial matter (e.g., O, N, S, and H).

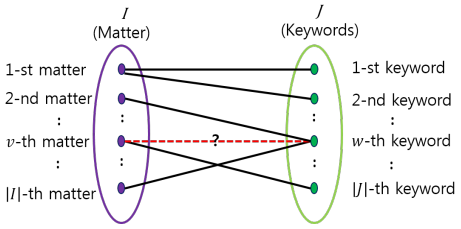

Figure 1: Bipartite graph with matter and keyword nodes

I \ J	1-st keyword	2-nd keyword	····	w-th keyword	····	|J|-th keyword
1-st matter	1	2		0		0
2-nd matter	0	0		1		
⋮						
v-th matter	0	0		0		3
⋮						
|I|-th matter	0	0		2		0

Figure 2: Example of an adjacency matrix R using the graph in Fig. 1

In the process of extracting keywords for constructing keyword nodes, first, we reduce each word to its root using stemming in each document composed of title and abstract. Then we use TF-IDF for each word and keep only the top 20 high TF-IDF valued keywords in each article, excluding the words of matter which cannot be in the keywords set of the bipartite graph. This allows us to select the important words in the paper as keywords which are likely to have a close and important relationship with the key matter in the title. As shown in Fig. 1, we construct a bipartite graph $G = (I+J, E)$ where I is the set of matter nodes, J is the set of keyword nodes, and E is the set of edges between nodes in I and J, which are formed when the two types of nodes appear in the same paper.

Its adjacency matrix $R \in \mathbb{R}^{|I|\times|J|}$ is defined as each element $r_{ij} = n$ where the matter of the i-th node and the keyword of the j-th node appear together in n different publications for $i \in I$ and $j \in J$. Fig. 2 shows the example of an adjacency matrix R using the graph in Fig. 1. To use the CF algorithms in our data set, we remove the matter which appears only once in the total publications to prevent the cold start problem [7]. In the case of keyword nodes, we select the keywords which appear more than 100 times in the total publications and remove trivial keywords which only consist of numbers. The target keywords are reduced to 'antiferromagnet' from 'antiferromagnetic' or 'antiferromagnetism' and 'superconduct' from 'superconductor(s),' 'superconducting,' or 'superconductivity' using stemming.

B. Link Prediction in the Bipartite Graph

CF is used for movie or product recommendation in several online services based on the user-item rating or purchase history. We consider the adjacency matrix R in the bipartite graph as a user-item matrix for CF algorithms. By using CF algorithms in the matrix R, we can predict formations of new links that are not contained in the link

set E of the bipartite graph G [6]. We consider the matter nodes as users and the keyword nodes as items.

For user-based CF, we need to calculate the similarity between pairs of matter. We use cosine-based similarity (sim) for all pairs of the matter in the set I [7]. In the next step, let $v \in I$ and $w \in J$ for which value of element r_{vw} in R is zero. The zero value in the matrix R represents that there is no link between the v-th matter and the w-th keyword. The following (1) is used when predicting the formation of new links with the user-based method [7]:

$$\acute{r}_{vw} = \bar{r}_v + \frac{\sum_{u \in U_m} (r_{uw} - \bar{r}_u) \cdot sim(v,u)}{\sum_{u \in U_m} |sim(v,u)|} \tag{1}$$

where $\bar{r}_v$ is the average value of non-zero elements in the v-th row in R, the set U_m is composed of the top-m most similar matter to the target v-th matter among the entire matter using the cosine-based similarity, $\bar{r}_u$ is the average value of non-zero elements in the row of the matter $u \in U_m$ in R, and $sim(v,u)$ represents the cosine similarity between v-th and u-th row in R. The predicted value $\acute{r}_{vw}$ represents how likely the link is formed in the future so a higher value indicates a higher probability of the link formation.

In the following section, we show that the appearance frequency of matter words in publications, which represents the popularity of matter, is the critical factor for the appearance frequency of matter in the future research. Therefore, we suggest considering the popularity of matter by summation of the number of times it appears in the publication data, for both perspectives of negative and positive effects on the formation of links in the future. The modified predicted value $\dot{r}_{vw}$ considering user-based method and matter popularity (user-based MP) is (2, 3):

$$s_{vw} = \bar{r}_v + \frac{\sum_{u \in U_m^*} (r_{uw} - \bar{r}_u) \cdot sim(v,u)}{\sum_{u \in U_m^*} |sim(v,u)|} \tag{2}$$

$$\dot{r}_{vw} = \log(\sum_{j \in J} r_{vj}) \times (s_{vw} + \alpha) \tag{3}$$

where U_m^* is the set composed of all elements in U_m, the top-m most similar matter to the target v-th matter, and additionally the v-th matter itself. Instead of U_m, we use U_m^* in (2) to consider the negative effect of matter popularity on the predicted value. For the negative effect, here is the explanation about the case when $u = v$ in the second term of (2). Note that the value of r_{vw} is zero in the matrix R and a larger $\bar{r}_v$ indicates that the v-th matter has more links, i.e. it is more popular. Therefore, if $\bar{r}_v$ is large, then the link formation between the v-th matter and the w-th keyword, which has not yet been formed, becomes a more rare event than the case when $\bar{r}_v$ is small. In other words, we can interpret the case when the $\bar{r}_v$ is large and r_{vw} is zero as the formation of the specific link is a rare event, because there is no link between the v-th matter and the w-th keyword even though the v-th matter has been researched a lot. The value of $r_{vw} - \bar{r}_v$ which is negative in (2) represents how rarely the link will be formed between the v-th matter node and w-th keyword node and the value decreases the predicted value considering the rareness of the link formation.

On the other hand, $\log(\sum_{j \in J} r_{vj})$ in (3) is the weighting value for the positive effect of matter popularity. The value of $\sum_{j \in J} r_{vj}$ is the summation of all values in the v-th row in matrix R and represents the popularity of the v-th matter in the publications. The more popular the matter is the more likely it is to have new links. The role of constant α in (3) is to make all negative predicted values of s_{vw} positive by positive parallel translation before they are weighted by the matter popularity. We sort the modified predicted values from user-based MP (3) in descending order. If the modified predicted value $\dot{r}_{vw}$ is high, the link has a higher probability to be formed in the future so the model recommends the links from the highest predicted valued link. Additionally we suggest an algorithm considering the matter popularity for the item-based algorithm (item-based MP) with the constant σ.

In the model-based algorithm for CF, we propose matrix factorization with matter popularity (MFMP). Let $P \in \mathbb{R}^{|I| \times K}$, $Q \in \mathbb{R}^{K \times |J|}$ be matrices with the parameter K of latent features number. The matrix factorization (MF) method is to find $\widehat{R} = PQ$ which is the approximated matrix to the true adjacency matrix R [8]. Let the i-th row in P be vector $\vec{p}_i$ and the j-th column in Q be vector $\vec{q}_j$. The process of finding $\widehat{R}$ is to use stochastic gradient descent for minimizing squared of error e_{ij}^2 in (4) with L2 regularization.

$$\begin{aligned} e_{ij} &= r_{ij} - \vec{q}_j^{\,\mathsf{T}} \vec{p}_i \\ \vec{q}_j &= \vec{q}_j + \gamma \cdot (e_{ij} \cdot \vec{p}_i - \lambda \cdot \vec{q}_j) \\ \vec{p}_i &= \vec{p}_i + \gamma \cdot (e_{ij} \cdot \vec{q}_j - \lambda \cdot \vec{q}_i) \end{aligned} \tag{4}$$

By calculating $\vec{q}_j$ and $\vec{p}_i$ iteratively using (4) for all $j \in J$ and $i \in I$ such that r_{ij} is not zero, we can get the optimized P and Q for getting the matrix $\widehat{R}$. In the MF method, the predicted value for the link between $v \in I$ and $w \in J$ is $\vec{q}_v^{\,\mathsf{T}} \vec{p}_w$. Considering the positive effect of matter popularity, the predicted value from MFMP is:

$$\dot{r}_{vw} = \log(\sum_{j \in J} r_{vj}) \times \vec{q}_v^{\,\mathsf{T}} \vec{p}_w \tag{5}$$

IV. Experiments

A. Datasets for the Recommendation System

We use 45,603 publications in PRB and PRL from 2004 to 2012 as a training set and 15,624 publications from 2013 to 2016 as a test set for retrospective study. By setting the test set as the more recent data than the training set we can evaluate the performance of the concept of predicting the future links formation. After preprocessing the data as mentioned in the Methodology section, we get a 2807×1782 matrix of R; the size of the matter set I is 2,807 and the size of the keyword set J is 1,782.

B. Distribution of Appearance Counts for Matter

We investigate the distribution of matter appearance counts in the titles and abstracts of the publications from 2000 to 2016. The plot in the Fig. 3 shows the log-log scale of cumulative distribution of the appearance counts of matter in the total publications following a straight line and it shows that most publications are concentrated on only a few most popular types of matter [9].

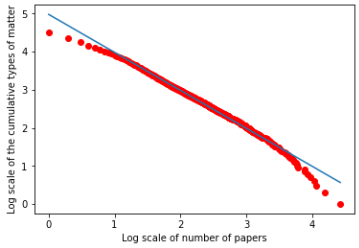

Figure 3: Log-log scale of cumulative distribution of appearance counts of matter in the total publications

C. Benchmark Algorithms for Comparison

We use the following methods for comparing the suggested algorithms: User-based MP, Item-based MP, and MFMP with the parameters m=10, α=2.4, σ=0.01, K=97, $\gamma = 0.0002$, and $\lambda = 0.01$.

1) User-based: Use the predicted value from (1).
2) Item-based: Use the predicted value with item similarity.
3) Preferential Attachment: $|\Gamma(x)| \times |\Gamma(y)|$ ($\Gamma(x)$: the set of neighbors of a node x). [6].
4) Matrix Factorization (MF): The predicted value is the element of the approximate matrix $\widehat{R}$ in the previous section.
5) Random: Recommend links randomly.

D. Investigation and Evaluation

We investigate two different aspects of hypotheses generation using link prediction.

1) We try to predict links in the range of the entire bipartite graph G. We compare the performance of user-based MP, item-based MP, and MFMP with five benchmark methods that we mentioned above. We evaluate each algorithm using the revised Global Receiver Operating Characteristic (GROC) curve [6], [10]. In the revised GROC curve, we evaluate the performance by increasing recommendation of links from the entire graph between matter and keywords without limiting the number of recommendations in each matter.
2) We try to predict links between matter and each specific stemmed keyword 'antiferromagnet' and 'superconduct'. The area under ROC curve (AUROC) is calculated for the performance of the suggested methods and benchmark algorithms.

V. Results and Discussions

A. Link Prediction for Matter and Keywords

For the first step, we compare the performance of link predictions in the entire graph G for suggested methods and benchmark methods. In Fig. 4, the item-based and item-based MP methods outperform the other methods, and the item-based MP is better than item-based in the range from about 100 to 250 recommendations.

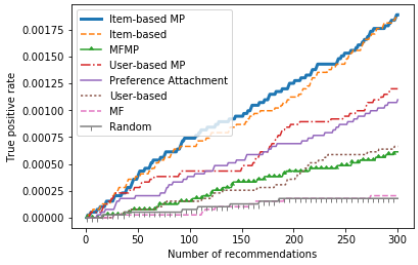

Figure 4: Revised GROC results for the algorithms

In the second step, we focus on the keywords 'antiferromagnet' and 'superconduct.' First, we perform experiments for the 'antiferromagnet' keyword. There are 2,360 zero elements among 2,807 elements in the column of the keyword 'antiferromagnet' in the matrix R. This represents that there are 2,360 possible new links. In the test set, there are 44 newly formed links for the keyword 'antiferromagnet.' The performance is the result of measuring how well each method recommends new links among 2,360 possible links for correctly predicting the 44 true links in the test set. For the second keyword 'superconduct,' there are 2,327 possible future links and 33 true new links in the test set.

Algorithms	AUROC
MF	0.5657
MFMP	0.6841
User-based	0.6754
User-based MP	**0.7755**
Item-based	0.7418
Item-based MP	**0.7614**
Preference Attachment	0.6837

Algorithms	AUROC
MF	0.5821
MFMP	0.6327
User-based	0.6962
User-based MP	**0.7350**
Item-based	0.5524
Item-based MP	0.6199
Preference Attachment	0.6303

Table I: AUROC for 'antiferromagnet' and 'superconduct'

Table I shows the AUROC value of each method for link prediction between matter and the each 'antiferromagnet' and 'superconduct' keyword. The bold values in the table are the largest ones or are not significantly different from the largest one at 98% confidence interval. The suggested methods, user-based MP and item-based MP, have better

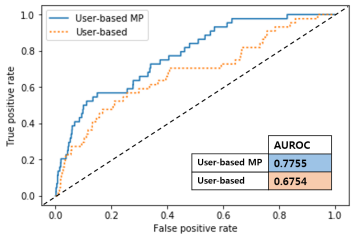

Figure 5: ROC curves of user-based MP and user-based method for 'antiferromagnet'

performance than the other methods for the 'antiferromagnet' and user-based MP is the best for the 'superconduct'. In Fig. 5, we can see the improved performance of user-based MP compared to the original user-based method for the 'antiferromagnet', with AUROC values of 0.7755 and 0.6754, respectively. The dashed line in the figure represents the performance of randomly selection.

Lastly, Table II shows detailed examples of link prediction results for the keyword 'antiferromagnet' within 100 recommendations. There is a total of 10 correct recommendations among 100. The predicted matter is always contained in the title but the keyword 'antiferromagnet' is either in the title or abstract. The matter and keywords are likely to have a close relationship because the matter in the title is the key matter for the paper and the key matter is related to the keywords.

Order in the list of recommendations	**Matter**	**Title or abstract**
6	Si	**Title**: **Antiferromagnetic** exchange interactions among dopant electrons in **Si** nanowires
8	Eu	**Title**: Effect of **Eu** magnetism on the electronic properties of the candidate Dirac material EuMnBi2. **Abstract**: Magnetic susceptibility measurements suggest **antiferromagnetic** (AFM) ordering of moments on divalent Eu ions near T-N = 22 K
15	FeSe	**Title**: Spin Ferroquadrupolar Order in the Nematic Phase of **FeSe**. **Abstract**: we find the FQ phase in close proximity to the columnar **antiferromagnet** commonly realized in iron-based superconductors.
22	Fe-doped	Omitted
23	Gd-doped	Omitted
28	SrTiO3	Omitted
43	Cu(001)	Omitted
52	Au	Omitted
55	Fe1-xTe	Omitted
98	Bi	Omitted

Table II: Detailed investigation of correctly prediction results for 'antiferromagnet' keyword within 100 recommendations

VI. Conclusions and Future Works

In this paper, we suggest a recommendation model for hypotheses generation in condensed matter physics. We construct the bipartite graph with matter words and keywords using publications and propose improved methods for predicting links in the graphs. From the results of the specific keywords, we confirm that our model can be applied to other various topics. Future works include applying our method to other various research topics or industrial topics using patent literature.

References

[1] P. O. Larsen and M. Von Ins, "The rate of growth in scientific publication and the decline in coverage provided by science citation index," *Scientometrics*, vol. 84, no. 3, pp. 575–603, 2010.

[2] G. Wallas, "The art of thought harcourt," *Bruce and Company, New York*, 1926.

[3] S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C. R. Pickering, A. Comer, J. N. Myers *et al.*, "Automated hypothesis generation based on mining scientific literature," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1877–1886.

[4] R. D. King, K. E. Whelan, F. M. Jones, P. G. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, no. 6971, pp. 247–252, 2004.

[5] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *journal of the Association for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[6] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decision Support Systems*, vol. 54, no. 2, pp. 880–890, 2013.

[7] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, p. 4, 2009.

[8] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[9] L. A. Adamic, "Zipf, power-laws, and pareto-a ranking tutorial," *Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html*, 2000.

[10] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 253–260.