Hypotheses generation
using link prediction in a bipartite graph.

Jung hun Kim
Graduate school of knowledge service engineering
KAIST
Daejeon, Korea
junghunkim@kaist.ac.kr

Abstract—The large volume of scientific publications is likely to have hidden knowledge which can be discovered dealing with the massive information altogether using a system which can cover the huge amount of data. We propose a method for generating hypotheses in the field of physics using the massive publications of physics journals. We convert the text data of titles and abstracts in publications to a bipartite graph extracting words of matter and keywords and suggest methods for predicting links between matter and keywords node. Those links generate hypotheses for research by suggesting the new possible relationship between some matter and phenomena or properties. Suggested methods have better performance than existed methods for link prediction in the bipartite graph. The experiments also show the detail application of our system for specific purpose: suggesting new matter which will be related with the concept of antiferromagnetism, superconductor and NMR spectroscopy.

Keywords-component; formatting; style; styling; insert (keywords)

I. INTRODUCTION

The volume of scientific publications is growing at an exponential rate [1, 13] which makes it impossible to keep up to date with all published papers. Automatic methods enabled by high performance computing and big data mining algorithms can generate insights at an aggregate level which would not otherwise be uncovered by looking at data silos independently. We suggest a system to analyze massive published literature for extracting knowledge and generating hypotheses. Previous work generated automated hypotheses in biology field using massive data from literature and experiments. We expand the field to physics especially the condensed matter physics which is the field of dealing with matter. The proposed system is helpful for researchers in the area of physics suggesting new research ideas based on relations between keywords and matter. New edges indicating the new relationships between matter and keywords are likely to contain new ideas which have not been thought or considered. By Wallas [27,28] generating new ideas depends on ‘Incubation’ which represents the subconscious without deliberate focus and ‘Illumination’ phase which represents a sudden flash of light. Because of the fact that there is vague process in generating ideas, Automated generation of hypotheses is a valuable tool with which human overcome the hardnes of generating ideas.

Publications from the Physical Review B (PRB) journal for the condensed matter physics and the Physical Review Letter (PRL) journal for general field of physics from 2004 to 2016 were used to analyze the proposed model and we divide the data in to two parts for train data set and test data set. We extract matter and keywords in the title and abstract of each article. We construct the bipartite network using two types of node, matter and keywords, and edges or links which are formed when the matter and keywords appear in the same article. Predicting the formation of new edge between nodes of matter and keywords indicates that the two entities will co-occur in future literature in this research area.

Condensed matter physics is one of the largest research fields in physics that deals with the physical properties of the phases of matter. In the condensed matter physics, the researchers seek to understand the behavior or property of matter in various conditions considering magnetization, electric field, mechanical stress and temperature change using special techniques or tools for analysis and find the application of the matter. Especially some special behaviors or properties have names like ‘superfluid’, ‘superconductivity’, ‘Bose-Einstein condensate (BCS)’ or ‘ferromagnetism’ which are considered as phenomena and some analyzing techniques also have names like ‘NMR Spectrometer’ or ‘X-ray Spectrometer’. For example, the relation between matter and phenomena is like ‘The matter ‘YBa$_2$Cu$_3$O$_7$’ has High-temperature superconductivity phenomenon’. The study of such phenomena in matter is an interesting research topic and those phenomena are normally important keywords in papers.

Antiferromagnetism is one of the magnetic properties in matter. This magnetic property means that the magnetic moments of atoms or molecules align in opposite direction of spins of electrons. The property is important to read elements of hard disk heads. Superconductivity is the one of the hottest research topics in the condensed matter because superconductivity is an interesting phenomenon of exactly zero electrical resistance and has many applications like MRI machines, magnetic fusion devices and maglev trains. Lastly the NMR Spectrometer is the technique for analyzing physical and chemical properties of atoms or molecules. It is
widely used for chemists and biochemists to investigate the properties of organic molecules. We focus on the antiferromagnetism, superconductivity and NMR Spectroscopy in the keyword nodes. Those words are stemmed to ‘antiferromagnet’ from ‘antiferromagnetic’ or ‘antiferromagnetism’, ‘superconduct’ from ‘superconductor(s)’, ‘superconducting’ or ‘superconductivity’ and ‘NMR’ respectively. Predicting link between matter and those nodes represents that we can predict new matter which will be revealed to have a relationship with the specific keywords.

The paper is organized as follow. Section II reviews the related work on recommendation algorithms, link prediction and hypotheses generation. Section III describes the suggested method including construction of the network and application of the improved collaborative filtering (CF) algorithms for the link prediction. Section IV shows the experiments for the analysis of the graph compared to other methods, explanation of how to evaluate the performances of the methods for the entire graph and subgraph including matter and the specific keyword node like ‘superconduct’, ‘antiferromagnet’ or ‘NMR’. Using the subgraph shows more practical usage of the method with detail investigation of real sentences in publications. Section V shows the result of the experiments and discussion with limitations of the proposed system. The last Section VI summarizes the proposed method and suggests future works.

II. RELATED WORK

A. Recommendation Algorithms

Recommendation algorithms are widely used for recommending items for users in the online services related with movie, song, Internet and products in markets like Netflix or Amazon corporation. CF is widely used technique in recommender systems and the main concept of it is predicting the future preferences of a target user using the known preferences of other users. There are two types of collaborative filtering: memory-based and model-based. Memory-based algorithms use the patterns of similar users to the target user. Therefore, the similarity between users need to be calculated using cosine similarity or Pearson correlation [10]. Considering the preferences of top 10 similar users to the target user the predicted preference values of the user for items are calculated and a system recommends items based on the values. The model-based algorithms normally use the matrix factorization (MF) technique. There are several ways for MF: The Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Probabilistic Matrix Factorization (PMF) and Non-Negative Matrix Factorization (NMF) [11]. Using the aforementioned techniques, the approximated matrix to the original matrix consisting of the preference of users is calculated and the predicted values are obtained from the calculated matrix. The CF using MF works well for finding the hidden structure behind the data [11].

B. Link Prediction in a Graph

The recommendation problem can be seen as a link prediction in a bipartite network. Link predictions is an active research area in computer science and normally the type of graph is an unipartite network such as the social network, webpages, and citation network. Liben-Nowelly [20] suggested the idea for link prediction in the co-authorship network for predicting future interactions between researchers using measurements for network topologies. L’u [23] suggested the methods for link prediction in a complex network using physical perspectives and approaches, such as the random-walk-based methods and the maximum likelihood methods. In the case of link prediction for the bipartite network there is previous work using CF algorithms, graph measures like common neighbors, preferential attachment and Katz [22] and graph kernel-based machine learning [7]. The link prediction performance tends to depend on the domain of the network data.

C. Hypotheses Generation

There were some trials for making systems for generating hypotheses for research using text mining of the scientific literature of Medline abstracts or algorithms for analysis of DNA data [2, 16, 24] in biology field. Spangler [2] constructed a system that can find the new protein kinases that phosphorylate the protein tumor suppressor p53 using graph-based diffusion of information. The system shows meaningful precision and recall in the retrospective study. In genetics, a sequence motif is a nucleotide pattern and by using algorithms of motif-finding method, Hu [16] analyzed the potential motif combination for hypotheses generation using DNA array data set and King et al.[24] applied a system to the determination of gene function using deletion mutants of yeast which competitive with human performance.

III. METHODOLOGY

A. Construction of the Bipartite Network

For construction of nodes in the bipartite graph we only consider the train data set and exclude the new nodes in the test data set. First we extract words of matter only in the titles for considering the key matter in each paper. In the titles, first we remove the special characters (e.g. ‘(’, ‘)’, ‘-’) in the words and then check whether they are composed of the list of all the
chemical elements in the periodic table (e.g. Li, Ne, Ca), names of special materials and some extra characters or notations. The following describes the text patterns used to extract words related to matter.

- There is matter which is composed of the list of the 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 ‘deta’, ‘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 matter which length is one because normally the chemical elements which length is one are very common elements so it is hard to have special meaning (e.g. O, N, S and H).

In the process of extraction of keywords, first we stem each word and then use the TF-IDF for the words in each document composed of title and abstract of each paper and keep only the top 20 high TF-IDF valued keywords in each article extracting the name of matter. This allows us to select the important words as keywords which are likely to have direct relationship with the key matter in the title. As shown in the 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, \( E \) is the set of edges between nodes in \( I \) and \( J \) which are formed when the two type of nodes appear in the same paper and its adjacency matrix \( R \in \mathbb{R}^{[I \times J]} \) is defined as each element \( r_{ij} = n \) where \( i \in I, j \in J \) and the matter of \( i \)-th node and the keyword of \( j \)-th node appear at the same time in \( n \) different publications in the training set. For using the CF algorithms in our data set we remove the matter which appear only once in the training set to prevent the cold start problem [6] which occurs when a new matter has less information for finding the similar matter in matter nodes. In the case of keyword nodes we select the keywords which appear more than 100 times in the total publications and remove some trivial keywords which only consist of numbers.

B. Link Prediction in the Bipartite Graph

The collaborative filtering (CF) is used for recommendation such as the movie recommendation considering the user-item rating as a matrix in Netflix and product recommendation to their customers in Amazon.com based on purchase history [25]. 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 new formation of links which is not contained in previous link set \( E \) of the bipartite graph \( G \) [5]. We consider the matter nodes as users and the keyword nodes as items. For memory based algorithm of User-based CF we need to calculate the similarity between matter. We use cosine-based similarity (1) for the all pairs of the matter in the set \( I \) [6]:

\[
\text{sim}(v_1, v_2) = \cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|_2 \|v_2\|_2} \tag{1}
\]

where \( v_1, v_2 \in I \) and \( v_1, v_2 \) are \( v \)-th and \( v \)-th row vector in \( R \) respectively. In the second step let \( v \in I \) and the \( w \in J \) for which value of element \( r_{v,w} \) 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. Prediction of the formation of new links with the User-based method uses the following (2) [6,7]:

\[
\hat{r}_{v,w} = \bar{r}_v + \frac{\sum_{u \in U}(r_{u,w}-\bar{r}_u) \times \text{sim}(v,u)}{\sum_{u \in U} |\text{sim}(v,u)|} \tag{2}
\]

where \( \bar{r}_v \) is the average value of non-zero elements in the \( v \)-th row in \( R \), the set \( U \) is composed of the most top 10 similar matter to the \( v \)-th matter and \( \bar{r}_u \) is the average value of non-zero elements in the row of the matter \( u \in U \) in \( R \). The predicted value \( \hat{r}_{v,w} \) represents how likely the link is formed in the future so higher value indicates the higher probability of formation.

In the following Section IV, we can see that the popularity of matter in the past is critical for the distribution of appearance counts of matter in the future. In the suggested algorithm, we consider popularity of matter from the appearance counting number in the training set for negative and positive effects. The modified predicted value \( \tilde{r}_{v,w} \) considering User-based method and matter popularity (User-based MP) is the following (2, 3):

\[
s_{v,w} = \bar{r}_v + \frac{\sum_{u \in U}(r_{u,w}-\bar{r}_u) \times \text{sim}(v,u)}{\sum_{u \in U} |\text{sim}(v,u)|} \tag{2}
\]

\[
\hat{r}_{v,w} = \log_{10}(\sum_{j \in J} r_{v,j}) \times (s_{v,w} + \alpha) \tag{3}
\]

Where \( U^* \) is the set composed of the most top 10 similar matter to the \( v \)-th matter and the \( v \)-th matter itself and \( \sum_{j \in J} r_{v,j} \) is the summation of the all values in the \( v \)-th row in matrix \( R \) which represents the popularity of \( v \)-th matter. The reason why we use \( U^* \) by adding the matter \( v \) itself to the set \( U \) is that for considering the negative effect of matter popularity. If the \( \bar{r}_v \) is large and the \( \bar{r}_w \) is zero then the link between the \( v \)-th matter and the \( w \)-th keyword represents a rarer event than the case that \( \bar{r}_w \) is small. The gap between \( r_{v,w} \) which is zero and \( \bar{r}_v \) represents that the negative effect of matter popularity. The \( \log_{10}(\sum_{j \in J} r_{v,j}) \) represents the positive effect of matter popularity. The higher the matter popularity is the more likely new links are formed. The \( \log_{10} \) alleviates the impact of the large value of matter popularity. The \( \alpha \) makes the negative values of \( s_{v,w} \) to positive. We sort the modified predicted values from User-based MP (3) in descending order. The higher modified predicted value \( \hat{r}_{v,w} \)
represents having higher probability for the link to be formed in the future so the system can recommend the links from the highest predicted valued link. Another algorithm for memory based algorithm is Item-based CF which is similar with concept of User-based CF except considering the similarity between items rather than users using (1) [5]. The formula is following:

\[
\hat{r}_{v,w} = \bar{r}_v + \frac{\sum_{x \in X}(r_{v,x} - \bar{r}_x) \cdot \text{sim}(v,x)}{\sum_{x \in X} \text{sim}(v,x)}
\]

(4)

where the set X is composed of the most 10 similar keywords to the w-th keyword and instead of using \(\bar{r}_w\) which is the average value of non-zero elements in the w-th column in \(R\) we use \(\bar{r}_v\) for considering the average value of v-th matter [5]. With the same perspective of the User-based algorithm, we suggested new algorithm considering the negative and positive effect of the matter popularity. The suggested formulas are following:

\[
s_{v,w} = \bar{r}_v + \frac{\sum_{x \in X}(r_{v,x} - \bar{r}_x) \cdot \text{sim}(v,x)}{\sum_{x \in X} \text{sim}(v,x)}
\]

(4)

\[
\hat{r}_{v,w} = \log_{10}(\sum_{x \in X} r_{v,x} \cdot s_{v,w}) \times (s_{v,w} + \sigma)
\]

(5)

where the \(X^*\) is the set composed of the w-th keyword and the elements in the set X. Equation (4) shows the predicted value considering the negative effect of matter popularity. The alleviated matter popularity for Item-based CF (Item-based MP) is considered in (5) with \(\sigma\) for making the negative values of \(s_{v,w}\) positive.

In the model-based algorithm for CF, we consider the Matrix Factorization with matter popularity(MFMP). Let the matrix \(P(\|l\| \times K)\) and \(Q(K \times \|l\|)\) where \(K\) is the parameter of latent features number. Matrix factorization(MF) method finds the \(\tilde{R} = PQ\) which is the approximated matrix to the adjacency matrix \(R\) [19]. Let the i-th row in \(P\) be vector \(p_i\) and the j-th column in \(Q\) is vector \(q_j\). Equations (4,5,6) are for obtaining the \(q_j\) and \(p_i\) which minimize the error \(e^2_{ij}\).

\[
e_{ij} = r_{ij} - q_j^T p_i
\]

(4)

\[
q_j = q_j + \gamma \cdot (e_{ij} \cdot p_i - \lambda \cdot q_j)
\]

(5)

\[
p_i = p_i + \gamma \cdot (e_{ij} \cdot q_j - \lambda \cdot p_i)
\]

(6)

By calculating \(q_j\) and \(p_i\) iteratively using (4,5,6) for all \(j \in J\) and \(i \in I\) such that \(r_{ij}\) is not zero we can get optimized \(P\) and \(Q\) for getting the matrix \(\tilde{R}\) and \(q_w^T p_w\) is the predicted value from MF method. Lastly we calculate the predicted value from MFMP by multiplying the alleviated popularity of matter to the value from MF as shown in the equation below:

\[
\hat{r}_{v,w} = \log_{10}(\sum_{x \in X} r_{v,x} \cdot q_w^T p_w)
\]

(7)

Equation (7) is the predicted value from MFMP method. Similar to User-based MP, we recommend the links by the predicted value \(\hat{r}_{v,w}\) in descending order.

IV. EXPERIMENTS

A. Distribution of Appearance Counts of Matter

We investigate the distribution of appearance counts of matter in the titles and abstracts of publications in PRB and RPL journals from the 2000 to 2016. The y-axis of Fig. 2 indicates the number of different types of matter and x-axis indicates the number of papers that a matter appears. This figure seems to be a power law distribution. For more detail investigation of the plot we modify the scale of the plot as in Fig. 3.

Fig. 3 shows the log scale of Fig. 2 for spreading the data points and we can see the straight line in the range of upper value of y-axis and spreading points in the low value of y-axis which is the evidence of the power law distribution.
Equations (4,5) are the cumulative distribution of power law distribution and the log-log scale of it.

\[ 1 - F(x) = P(X \geq x) = \frac{C}{1-\gamma}x^{1-\gamma} \]  

\[ \log P = C + (1 - \gamma) \log x \]  

Fig. 4 presents a straight line and the fitted straight line has the slope of -0.9975. It means the distribution is similar with the cumulative distribution of log-log scale of power law distribution (5) with \( \gamma = 1.9975 \) [14]. From the Fig. 2, Fig. 3 and Fig. 4 we can say that the distribution of appearance counts of matter follows the power law distribution.

In economics, distributions of the upper income tail and wealth of Hungarian medieval society follow the Pareto’s law or power law distribution [3, 15]. It represents that the most of the wealth is concentrated on a small group of rich people. A power law form for the wealth distribution results in which the rich get richer and the poor get poorer [26]. The power law in the distribution of the appearance counts of matter in the total published papers shows that most publications are concentrated on a few most popular matter and we can expect that the trend of it has continued and will continue like well known ‘the rich get richer’ trend in wealth. In the list of most frequently appearing matter is normally the single atom (e.g. Si, Fe, Co) and right after that the popular matter like ‘GaAs’, ‘diamond’ and ‘graphite’ are following.

B. Datasets for the Recommendation System

We use 45,603 publications during 2004~2012 in PRB and PRL as a training set and 15,624 publications during 2013~2016 as a test set for retrospective study. By setting the test set as the more recent data than the train set we can evaluate the performance for the concept of predicting future using data from the past. After the preprocessing of the data as mentioned in Section III we get a 2807×1782 matrix of \( R \).

C. Benchmark Algorithms for Comparison

We used the following methods for comparison with the suggested algorithms: User-based MP and MFMP with the parameters \( \alpha = 27.13, K=97, \gamma = 0.0002 \) and \( \lambda = 0.01 \)

1. User-based: Simply using the predicted value from (2) [17].
2. Item-based: Simply using the predicted value from (4) [5].
3. Preferential Attachment: For a node \( x \), we define \( \Gamma(x) \) as the set of neighbors of \( x \). A Preferential Attachment \( \Gamma(x) \cdot \Gamma(y) \) that recommends links according to the matter’s popularity and keyword popularity [22].
4. Matrix Factorization(MF): It is closely related to the singular value decomposition (SVD). The predicted value is obtained from (4,5,6) [18,7].
5. Random: Randomly choose the links for recommendation.

D. Investigation and Evaluation

We investigate into two different aspects for hypotheses generations 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 5 benchmark methods that we mentioned above. We evaluate each algorithm using modified global receiver operating characteristic (GROC) curve which is little bit different from ROC or revised ROC curve [7, 23]. In a GROC curve [23] rather than evaluating performance by recommending top-k links for each matter, we evaluate the performance by recommending links in the total possible new links between matter and keywords without limitation of recommended number in each matter. Therefore, the number of recommended links in each matter does not need to be unequal. In the revised GROC [7], x-axis is the number of recommendations rather than false positive rate which is used in typical GROC. The two variables are highly correlated so that it doesn’t change the shape of the curve much and the revised curve is more straightforward for understanding the performance of the methods. After that we plot the precision rate (8) for each method by increasing the recommendation number.

\[
\text{Precision} = \frac{\text{Number of recommended links that match with future links}}{\text{Total number of recommended links}}
\]  

In the practical aspect of the system the precision is more important than recall because the experiments for proving the recommendations from the system make a false positive costly. Precision represents how much the recommendations from each method is truthful.

2. We try to predict links between matter and the specific keyword ‘superconduct’. Because of the importance of the precision we plot the precision by increasing the number of recommendation. We evaluate performance at 10 and 50 number of
recommendations calculating the precision, recall (9) and \( F_{0.5} \)-measure (10); \( F_{0.5} \)-measure is suitable for emphasizing the precision.

\[
\text{Recall} = \frac{\text{Number of recommended links that match with future links}}{\text{Total number of links formed in the future}} \tag{9}
\]

\[
F_{0.5} - \text{measure} = \frac{[1+0.5^2] \times \text{Precision} \times \text{Recall}}{0.5^2 \times \text{Precision} \times \text{Recall}} \tag{10}
\]

The reason why we investigate the performance at the small number of recommendations, 10 and 50, is that the precision rate is decreasing by increasing the number of recommendations. In addition, the high precision is important rather than high recall for the practical aspect of the system even though the number of recommended links is small. Small number of correctly predicted link is valuable enough for researchers. The area under GROC curve (AUGROC) is calculated for checking whether the concept of our suggested recommendation system is valuable by evaluating the performance from minimum of the threshold for the number of recommendations to the maximum of it but AUGROC is less important evaluation for the practical aspect. After those evaluations we investigate more detail example about the predicted links between matter and the keyword ‘superconduct’ using real sentences which include the two words of nodes in the publications.

V. RESULTS AND DISCUSSIONS

A. Link prediction for matter and keywords

For the first step, as mentioned in Section IV we compare the performance of suggested models with benchmark methods for the entire graph \( G \). Fig. 5 shows the revised GROC curves of 300 recommendations.

![Figure 5. Modified GROC results for the algorithms.](image)

The graph shows the performance of the link prediction between matter and the ‘antiferromagnet’ keyword using real sentences which include the two words of nodes in the publications.

For the prediction in the entire graph, the Item-based and Item-based MP method shows the outperformance comparing other methods and the Item-based MP is better than Item-based in range from 100 to 250 recommendations. User-based MP is following except in the range about 150 recommendation numbers where the Preference Attachment shows better performance. Suggested methods of Item-based MP, User-based MP and MFMP normally shows better performance than original methods of Item-based, User-based and MF, respectively. From the Fig. 5 we can see the CF methods and Preference Attachment method works well for the prediction of links in the graph comparing the performance of the randomly recommendation. Interestingly the memory-based algorithms, Item-based and User-based, are better than model-based algorithms, MF and MFMP.

![Figure 6. Precision rate for the algorithms.](image)

Fig. 6 shows the precision rate for 300 recommendations. Normally the precision rate is the highest in the first few recommendations within 50 recommendations. After the point of the highest precision it decreases and converge to some precision rate by increasing number of recommendations because each method recommends links by the descending order of the predicted value which represents the probability of formation of link. The Item-based, Item-based MP methods and User-based MP shows high precision rate from the very first number of recommendations and Item-based and Item-based MP shows outperformance in the most of range.

In the second step as we mentioned in Section IV, instead of the entire graph \( G \), we focus on the keyword ‘antiferromagnet’, ‘superconduct’ and ‘NMR’. The result of the experiments shows the performance of the link prediction between matter and the specific keywords.

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

Table 1 shows the AUROC value of each method for link prediction between matter and the ‘antiferromagnet’ keyword. The bold values are largest one and not significantly different from the largest one at 95% confidence interval. The suggested methods considering matter popularity, User-
based MP and Item-based MP, have better performance than the other methods. MFMP shows better performance than MP method.

![Figure 7. ROC curves of User-based MP and User-based method for 'antiferromagnet.'](image)

In the Fig. 7, we can see obviously the improved performance of the method of User-based MP comparing the original User-based method.

**TABLE II. DETAIL INVESTIGATION OF PREDICTION RESULT FOR ‘ANTIFERROMAGNET’ KEYWORD WITHIN 100 RECOMMENDATIONS**

| Matter | Title or abstract |
|--------|-------------------|
| 6      | **Si**<br>Title: Antiferromagnetic exchange interactions among dopant electrons in Si nanowires. | |
| 8      | **Eu**<br>Title: Effect of Eu magnetism on the electronic properties of the candidate Dirac material EuMnBi2. <br>Abstract: Magnetic susceptibility measurements suggest antiferromagnetic (AFM) ordering of moments on divalent Eu ions near T_N = 22 K. | |
| 15     | **FeSe**<br>Title: Spin Ferroquadrupolar Order in the Nematic Phase of FeSe. <br>Abstract: we find the FQ phase in close proximity to the columnar antiferromagnet commonly realized in iron-based superconductors. | |
| 22     | **Fe-doped**<br>Title: Magnetic phase transition in Fe-doped topological insulator Bi2Se3 <br>Abstract: For higher Fe concentration, >1.7 at. %, Bi2Se3 prefers the antiferromagnetic phase mediated by the superexchange interaction. | |
| 23     | **Gd-doped**<br>Title: Systematic study of the exchange interactions in Gd-doped GaN containing N interstitials, O interstitials, or Ga vacancies. <br>Abstract: The exchange interactions between N interstitials (Ni) and Ni with Gd are found to be short-ranged and mainly antiferromagnetic. | |
| 28     | **SrTiO3**<br>Title: First-principles study of magnetic frustration in FeSe epitaxial films on SrTiO3 <br>Abstract: For monolayer FeSe epitaxial film, the combined effect of electron doping and phonon vibrations readily leads to magnetic frustration between the collinear antiferromagnetic state and the checkerboard antiferromagnetic N'ee'l state. | |

**Table 2 shows the AUROC values for the link prediction of the keyword ‘antiferromagnet’.** The suggested methods of MFMP and User-based MP and the Preference Attachment method show better performance than others. The Fig. 7 shows the ROC curves of the User-based MP which shows the best performance in Table 2 and User-based method for the ‘antiferromagnet’ keyword. The suggested method of User-based MP shows the much better result than the existed method of User-based.

**TABLE III. AUROC FOR ‘SUPERCONDUCT’**

| Algorithms | AUROC |
|------------|-------|
| MFMP       | 0.6515 |
| User-based | 0.6771 |
| Preference Attachment | 0.6467 |
| User-based | 0.6406 |
| Item-based | 0.6022 |
| Item-based MP | 0.6531 |
| MF          | 0.6200 |

Table 1 shows the AUROC value of each method for link prediction between matter and the ‘superconduct’ keyword and the bold values are not significantly different from the largest ones at 96% confidence interval. The suggested methods considering matter popularity, MFMP, User-based MP and Item-based MP, have better performance than the other methods especially the User-based MP shows the best performance among the entire methods.

**TABLE IV. DETAIL INVESTIGATION OF PREDICTION RESULT FOR ‘SUPERCONDUCT’ KEYWORD WITHIN 100 RECOMMENDATIONS**

| Recommendation number | Matter | Title or abstract |
|-----------------------|--------|-------------------|
| 35                    | Ni-doped | Title: Magnonlike Dispersion of Spin Resonance in Ni-doped BaFe2As2. <br>Abstract: Inelastic neutron scattering measurements on Ba(Fe0.963Ni0.037)2As2 manifest a neutron spin resonance in the superconducting state with anisotropic dispersion within the Fe layer. | |
| 55                    | MoS2   | Title: Strongly enhanced superconductivity in doped monolayer MoS2 by strain. | |
| 65                    | TiO2   | Title: Role of double TiO2 layers at the interface of FeSe/SrTiO3 superconductors. | |

Table 2 shows the detail investigation of link prediction result. The 35-th predicted link is formed between the matter ‘Ni-doped’ in the title and the keyword ‘superconducting’ in the
abstract. The other 55-th and 65-th predicted links are formed between matter MoS2 and TiO2 and keywords ‘superconductivity’ and ‘superconductors’ respectively which are all in titles.

In the Fig. 8 the Item-based MP shows the best performance within about 50 recommendation numbers but the User-based MP shows the best performance in the range above about 50. Table 4 shows the detail investigation of the contents for the correctly predicted links. Within 100 recommendations there are 9 correctly predicted links.

| Table V. AUROC for ‘NMR’ |
|---------------------------|
| Algorithms | AUROC |
| MFMP       | 0.7284 |
| User-based MP | 0.7414 |
| Preference Attachment | 0.7291 |
| User-based | 0.5899 |
| Item-based | 0.7016 |
| Item-based MP | 0.7277 |
| MF           | 0.5834 |

Table 3 shows the performance for the ‘NMR’ keyword and the User-based MP shows the best performance.

There are some limitations in this system. First thing is in the data set we use for this system. We consider only two journals, PRL and PRB, so even though there is no link in the train set the link exists in the other publications during the same period of our train set. Same like in the train set, in the case of test set even though there is no link in the test set there can be the link in other publications during the same period of our test set. Second as we can see in Fig. 6 and Fig. 7 the method of the best performance is different depending on the purpose like for the entire graph or specific keyword. In the last even though the system can correctly predict the new links there is possibility that appearance of keyword and matter in the same paper doesn’t always mean they are directly related with each other.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we suggest the new recommendation system for hypotheses generation for condensed matter physics. We convert the publication data into the bipartite graph using matter words and keywords and using user-based MP and MFMP methods we suggest new predicted links in the future graph. We can get better performance in our suggested method in various purpose like for the entire graph and subgraph containing the ‘antiferromagnet’ keyword node.

In the future works, we will apply our system for expanding data set including publications in other physics journals. Also, In addition to finding links between matter and the keyword ‘antiferromagnet’, we can try to apply our method to the other subgraph for specific purpose such as the keywords like ‘superfluid’, ‘BCS’ and ‘ferromagnetism’ which are important concept in the condensed matter physics.

ACKNOWLEDGMENT

REFERENCES

[1] Larsen, Peder Olesen, and Markus Van Ins. "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index." Scientometrics 84.3 (2010): 575-603.
[2] Spangler, Scott, et al. "Automated hypothesis generation based on mining scientific literature." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.
[3] Zhu, Fei, et al. "Biomedical text mining and its applications in cancer research." Journal of biomedical informatics 46.2 (2013): 200-211.
[4] Clementi, Fabio, and Mauro Gallegati. "Pareto’s law of income distribution: Evidence for Germany, the United Kingdom, and the United States." Econophysics of wealth distributions. Springer Milan, 2005. 3-14.
[5] Sarwar, Badrul, et al. "Item-based collaborative filtering recommendation algorithms." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.
[6] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." Advances in artificial intelligence 2009 (2009): 4.
[7] Li, Xin, and Hsinchun Chen. "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach." Decision Support Systems 54.2 (2013): 880-890.
[8] Resnick, Paul, et al. "GroupLens: an open architecture for collaborative filtering of netnews." Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994.
[9] Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "Review Article a Survey of Collaborative Filtering Techniques" published in Hindawi Publishing Corporation. Advances in Artificial Intelligence 2009.
[10] Resnick, Paul, et al. "GroupLens: an open architecture for collaborative filtering of netnews." Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994.

[11] Bokde, Dheeraj, Sheetal Girase, and Debajyoti Mukhopadhyay. "Matrix factorization model in collaborative filtering algorithms: A survey." Procedia Computer Science 49 (2015): 136-146.

[12] Srinivasan, Padmini. "Text mining: generating hypotheses from MEDLINE." Journal of the Association for Information Science and Technology 55.5 (2004): 396-413.

[14] Adamic, Lada A. "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).

[15] Hegyi, Geza, Zoltan Neda, and Maria Augusta Santos. "Wealth distribution and Pareto's law in the Hungarian medieval society." Physica A: Statistical Mechanics and its Applications 380 (2007): 271-277.

[16] Hu, Yuh-Jyh, et al. "Combinatorial motif analysis and hypothesis generation on a genomic scale." Bioinformatics 16.3 (2000): 222-232.

[17] Pazzani, Michael J. "A framework for collaborative, content-based and demographic filtering." Artificial intelligence review 13.5-6 (1999): 393-408. Artificial Intelligence Review, 13 (1999), pp. 393-408

[18] Ott, Patrick. Incremental matrix factorization for collaborative filtering. Hochsch. Anhalt. Presse-und Öffentlichkeitsarbeit, 2008.

[19] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer 42.8 (2009).

[20] Liben-Nowell, David, and Jon Kleinberg. "The link-prediction problem for social networks." journal of the Association for Information Science and Technology 58.7 (2007): 1019-1031.

[22] Chen, Hsinchun, Xin Li, and Zan Huang. "Link prediction approach to collaborative filtering." Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on. IEEE, 2005.

[23] Schein, Andrew I., et al. "Methods and metrics for cold-start recommendations." Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2002.

[24] King, Ross D., et al. "Functional genomic hypothesis generation and experimentation by a robot scientist." Nature 427.6971 (2004): 247.

[25] Ekstrand, Michael D., John T. Riedl, and Joseph A. Konstan. "Collaborative filtering recommender systems." Foundations and Trends® in Human–Computer Interaction 4.2 (2011): 81-173.

[26] Ispolatov, Slava, Paul L. Krapivsky, and Sidney Redner. "Wealth distributions in asset exchange models." The European Physical Journal B-Condensed Matter and Complex Systems 2.2 (1998): 267-276.

[27] G. Wallas, The Art of Thought., Harcourt Brace, New York, 1926.

[28] Herring, Scarlett R., Brett R. Jones, and Brian P. Bailey. "Idea generation techniques among creative professionals." System Sciences, 2009. HICSS’09. 42nd Hawaii International Conference on. IEEE, 2009.