Prediction of Company’s Trend based on Publication Statistics and Sentiment Analysis

Fumiyo Fukumoto¹, Yoshimi Suzuki¹, Akihiro Nonaka¹ and Karman Chan²

¹Graduate Faculty of Interdisciplinary Research, Univ. of Yamanashi, 4-3-11, Takeda, Kofu, Yamanashi, Japan
²Business Strategy and Development Center, HIJ Innovation Institute Inc., 2-10-2, Fujimi, Chiyoda-ku, Tokyo, Japan

Keywords: Publication Statistics, Sentiment Analysis, Prediction, Company’s Trend.

Abstract: This paper presents a method for predicting company’s trend on research and technological innovation/development (R&D) in business area. We used three types of data collections, i.e., scientific papers, open patents, and newspaper articles to estimate temporal changes of trends on company’s business area. We used frequency counts on scientific papers and open patents to be published in time series. For news articles, we applied sentiment analysis to extract positive news reports related to the company’s business areas, and count their frequencies. For each business area, we then created temporal changes based on these frequency statistics. For each business area, we clustered these temporal changes. Finally, we estimated prediction models for each cluster. The results show that the model obtained by combining three data is effective to predict company’s future trends, especially the results show that SP clustering contributes overall performance.

1 INTRODUCTION

With the exponential growth of industries, an enormous body of companies provide jobs with various business areas, e.g., IT engineer, Electronic communication, and medical science. Hence, given the limited time, people, especially students have to go on job hunting. Ideally, many would be more interested in getting a job that matches their expertise of study, and companies having high prospect in those fields in the future. However, it is difficult to make a decision that whether a company would really make an expertise and investment in the future before working at that company. Therefore, it is beneficial for automatically predicting company’s trend on R&D in various business fields.

In the context of high impact academic papers prediction, citation-based analysis is often used (McGovern et al., 2003; Bethard and Jurafsky, 2010). Adams showed that the number of citations is used to quantify paper’s attention (Adams, 2005). McNamara et al. proposed a method for predicting paper’s future impact by using topological features extracted from citation network (McNamara et al., 2013). In addition to topological features, Davletov et al. predicted high impact academic paper based on temporal features of citations (Davletov et al., 2014). There are a few academic paper prediction method used on textual features (Kogan et al., 2009; Joshi et al., 2010; Yagatama et al., 2011), while much of the previous work on paper prediction used mainly citation statistics (Shi et al., 2010; Yan et al., 2012). Koppel et al. attempted to classify news stories about a company according to its apparent impact on the performance of the company’s stock (Koppel and Shrimberg, 2004). Milea et al. proposed a method of predicting the movement of the MSCI EURO index based on European Central Bank statements by learning and extracting fuzzy grammars from the text of the ECB statements (Milea et al., 2010). Lavrenko et al. presented a unique approach to identifying news stories that influence the behavior of financial markets. They identified trends in time series using piecewise linear fitting and assigned labels to the trends according to an automated binning procedure (Lavrenko et al., 2000). However, most of the above approaches focused on company’s financial impact.

In this paper, we focus on company’s R&D in some business areas (henceforth referred to categories), and propose a method for predicting their trend. We used frequency counts on scientific papers and open patents to be published in time series to obtain temporal changes for categories in the company. Moreover, we used frequency counts on positive news reports, e.g., new product sales, a success of new technologies, and improvement of corporate results, and
obtained temporal changes for a specific category in the company. To do this, we applied sentiment analysis to the newspaper articles to extract positive news reports related to the companies, and created temporal changes. For each category, we clustered company’s changes. Similar to Davletov et al.’s method, we used a regression model, i.e., for each cluster, model prediction is estimated as a regression problem where the objective is to predict company’s future trend in a specific category.

2 FRAMEWORK OF THE SYSTEM

We used three types of Japanese data, i.e., scientific papers, patent documents, and news articles to predict future trend of a company with a specific category. Figure 1 illustrates our system. It consists of two procedures, (1) Learning prediction model, and (2) Prediction. Learning prediction model consists of four steps: (1-1) creation of temporal series, (1-2) clustering temporal series, (1-3) impact estimation by citation network, and (1-4) construction of a prediction model. The input of the procedure is a company name(i), its category(m), e.g. communication engineering, and a time period(t) which we want to predict. The output is a plotted figure represented by a polynomial regression(degree equals to three) where x-axis is a time period and y-axis refers to a value indicated future trend on R&D.

2.1 Learning Prediction Model

The company’s future prospect is estimated by using patterns for each category as time series, and citation network. The pattern is derived from the frequencies on scientific papers, open patents to be published, and positive news reports related to the company.

2.1.1 Creation of Temporal Series

For each category in each company c^m, where i is a company and m is a category, we count the cumulative number of times for three data: scientific papers, open patents, and positive news reports related to c^m within a time period,1 and create three temporal series, scientific papers (Tpa^m), patents (Tpt^m), and news reports (Tnr^m). News articles are not assigned categories, while scientific papers and open patents are classified into categories. Therefore, we assigned categories to the news articles by using open patents. More precisely, we collected open patents for each category, and concatenated them into one document. We applied a simple weighting method, tf*idf scheme (Salton and Buckley, 1990) for term weighting. We used a noun/compound noun word as a term. For each category, we selected the number of topmost s terms according to tf*idf values as a feature. For each news article, we count the number of terms, and classified each article into categories whose number of the features is largest.

To extract positive news reports, we applied sentiment analysis. For each category m, all news reports including company name are extracted. These news reports were parsed by the syntactic analyzer CaboCha (Kudo and Matsumoto, 2003), and all the dependency triples (rel,x,y) are extracted. Here, x refers to a noun/compound noun word related to the company name, y shows a verb or an adjective word. rel denotes a grammatical relationship between x and y. We classified rel into 9 types of Japanese particle. “g(a)ha”, “wo”, “ni”, “he”, “to”, “de”, “yori”, “kara”, and “made”. For instance, from the sentence including company name “Toyota-no-uriage(Sales of Toyota) ha koutixyoudatta(are good).” (Toyota Motor corporation’s sales were good.), we can obtain the dependency triplet, (ha, Toyota-no-uriage, koutixyoudatta). The triplet shows positive opinion.

We regarded each of the extracted dependency triplet as positive/negative if y in the triplet (rel,x,y) is classified into positive/negative classes in the Japanese sentiment polarity dictionary (Kobayashi et al., 2005). However, the dictionary makes it nearly impossible to cover all of the words in the news article corpus. For unknown verb or adjective words that were extracted from the news article corpus, but did not appear in any of the dictionary classes, we classified them into positive or negative class by using a topic model. Topic models such as probabilis-

---

1We set a time period to six months
2As a result of manual evaluation, we set s to 20,000.
tic latent semantic indexing (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are based on the idea that documents are mixtures of topics, where each topic is captured by a distribution over words. The topic probabilities provide an explicit low-dimensional representation of a document. They have been successfully used in many domains such as text modeling and collaborative filtering (Li et al., 2013). We applied LDA based on Gibbs sampling to a news article corpus consisted of triples, and classified unknown words into positive/negative classes.

Figure 2 illustrates the result obtained by LDA. The sentence marked with box includes unknown verb words, such as pick up and terminate, “positive” and “negative” refer to a sentence including positive and negative words appeared in the sentiment polarity dictionary. “Topic id” indicates id number assigned to each cluster/set. We need to estimate the number of topics k for the result obtained by LDA. As shown in Figure 2, the result can be regarded as a clustering result: each element of the clusters is positive/negative news reports according to the sentiment polarity dictionary, or unknown words. Therefore, we estimated the number of topics (clusters) k by using Entropy measure given by:

\[ E = -\frac{1}{\log k} \sum_{j}^{k} \frac{N_j}{N} \sum_{i} P(A_i, C_j) \log P(A_i, C_j). \] (1)

k’ in Eq. (1) refers to the number of clusters. P(A_i, C_j) is a probability that the elements of the cluster C_j assigned to the correct class A_i. N denotes the total number of elements and N_j shows the total number of elements assigned to the cluster C_j. The value of E ranges from 0 to 1, and the smaller value of E indicates better result. We chose the parameter k’ whose value of E is smallest. For each cluster, if the number of positive news reports is larger than those of negative ones, we regarded a triplet including unknown word in the cluster as positive and vice versa. For example, "picked up" in the Topic id1 cluster shown in Figure 2 is regarded to a positive as the number of positive and negative were two and zero, respectively.

We used the result of triples to classify test news reports. Like much previous work on sentiment analysis based on supervised machine learning techniques (Turney, 2002), or corpus-based statistics, we used Support Vector Machines (SVMs) to annotate automatically (Joachims, 1998). Each training news report is represented by a vector. Each dimension of a vector is positive/negative triples appeared in the news report, and the value of each dimension is a frequency of the triplet in the news report. Each test news report is also represented by a vector. Each dimension of a vector is a triplet appeared in the news report. We used pairwise classification. As a result, for each e_m^i, we count the cumulative number of times for positive news reports, and create a temporal series, T_{nr}^m. Finally, we added these three temporal series given by Eq. (2).

\[ T_{i}^{m} = \alpha T_{pa}^{m} + \beta T_{pt}^{m} + \gamma T_{nr}^{m}. \] (2)

T_{i}^{m} in Eq. (2) refers to the temporal series with category m of the company i. \( \alpha, \beta, \text{ and } \gamma \) are parameters indicating the degree of importance.

2.1.2 Clustering Temporal Series

We note that temporal series T_{i}^{m} is created for each category of the company. The result is a large number of temporal series, and some of them are very close with each other. We then applied spectral(SP) clustering technique (Ng et al., 2002) to classify them into some classes. Similar to other clustering algorithms, the SP clustering takes as input a matrix formed from a pairwise similarity function over a set of data points. For each category, we applied SP clustering to the matrix D_{nr}^m. Given a set of points \( C^m = \{ c_{m}^1, \cdots , c_{m}^{n} \} \) where \( c_{m}^i \) denotes company i in the category m, the algorithm is as follows:

1. Form a distance matrix D_{nr}^m \in R^{2} between all the number of n companies that are used in the training. The distance matrix for the category m, is given by D_{nr}^{m} = \sum | T_{ij}^{m} - T_{ji}^{m} |. D_{nr}^{m} indicates distance, L1-norm between company \( i \) and \( j \) in the category m. \( T_{ij} \) and \( T_{ji} \) refer to the temporal series of the company i and j at the time t.
2. D is transformed to an affinity matrix A_{ij}.

\[ A_{ij} = \begin{cases} \exp(-\frac{D_{nr}^{m2}}{\sigma^2}), & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \]

\( \sigma^2 \) is a parameter and controls the rate at which affinity drops off with distance.
3. The matrix \( L = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \) is created. \( D \) is a diagonal matrix whose \((i,j)\) element is the sum of \( A \)'s \( i\)-th row.

4. The eigenvectors and eigenvalues of \( L \) are calculated, and a new matrix is created from the vectors associated with the number of \( l \) largest eigenvalues.

5. Each item now has a vector of \( l \) coordinates in the transformed space. These vectors are normalized to unit length.

6. \( K\)-means is applied to \( S \) in the \( L\)-dimensional space.

### 2.1.3 Impact Estimation by Citation Network

In addition to temporal series, we estimated impact of open patents. We used the Markov Random Walk (MRW) model to ranking open patents related to \( c_i^m \). For each category, we created a graph \( G^m = (C^m, E^m) \) that reflects the relationships between companies in \( C^m \). \( C^m \) refers to a set of companies in the category \( m \). Each \( e_{ij}^m \in E^m \) is represented by a vector whose dimension of a vector corresponds to each company. Each element of a dimension is a citation count of the company. \( E^m \) is a set of edges, which is a subset of \( C^m \times C^m \). Each edge \( e_{ij}^m \in E^m \) is associated with an affinity weight. The weight is computed using the standard cosine measure between two companies. We applied MRW model for each category.

The transition probability from \( e_{ij}^m \) to \( e_{jk}^m \) is then defined by normalizing the corresponding affinity weight \( p(i \rightarrow j) = \frac{\text{aw}(i \rightarrow j)}{\sum_{l \neq m} \text{aw}(i \rightarrow l)} \). If \( \sum_{i} \text{aw} \neq 0 \), otherwise, 0. We used the row-normalized matrix \( U = (U_{ij})_{c_{m} \times c_{m}} \) to describe \( G^m \) with each company corresponding to the transition probability, where \( U_{ij} = p(i \rightarrow j) \). To make \( U \) a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to \( \frac{1}{c_{m}} \). The matrix form of the saliency score \( \text{Score}(c_i^m) \) can be formulated in a recursive form as in the MRW model: \( \tilde{\lambda} = \mu U^T \tilde{\lambda} + \frac{1}{|C^m|} \tilde{\epsilon} \), where \( \tilde{\lambda} = [\text{Score}(c_i^m)]_{i=1}^{c_{m}} \) is a vector of saliency score for the companies, \( \tilde{\epsilon} \) is a column vector with all elements equal to 1. \( \mu \) is a damping factor. We set \( \mu \) to 0.85, as in the PageRank (Brin and Page, 1998). The above process can be considered as a Markov chain by taking the companies as the states and the final transition matrix is given by Eq. (3), and each score of the companies is obtained by the principle eigenvector of the new transition matrix \( M \).

\[
M = \mu U^T + \left(1 - \mu \right) \frac{1}{|C^m|} \tilde{\epsilon} \tilde{\epsilon}^T. \tag{3}
\]

\( \tilde{\epsilon} \) in Eq. (3) is a column vector with all elements equal to 1. The principal eigenvector in the MRW model is obtained by the power method and inverse iteration method. For implementation, we used the Eigen library.\(^3\) We chose a vector with the largest eigenvalues. We normalized a vector, and obtained company’s rank scores.

### 2.1.4 Prediction Model

For each cluster obtained by SP clustering, we estimated prediction model for a category \( m \) of the company \( i \) at a time period \( t \), \( RD_i^m(t) \) which is given by Eq. (4).

\[
RD_i^m(t) = g_i^m(t) + \lambda R_i^m \tag{4}
\]

where \( g_i^m(t) \) in Eq. (4) is a prediction model obtained by using temporal series in the cluster \( s_i \) assigning the company \( i \) for the category \( m \), and \( R_i^m \) refers to the ranking score of \( i \) obtained by MRW. \( \lambda \) is a weighting parameter.

### 2.2 Prediction of Trend

Given a specific category \( m \) of the company \( i \) which we want to predict, we first calculate the minimum distance between its mean of the cluster \( \mu_j \) assigned to the category \( m \) and the temporal series of the company \( i \) by using Eq. (5).

\[
s_i = \arg \min \left| \left| T_i^m - \mu_j \right| \right|. \tag{5}
\]

For the rest of the prediction timeline, the chosen cluster’s polynomial is used for prediction.

### 3 EXPERIMENTS

We used scientific paper publications and open patents provided by JST,\(^4\) and Mainichi Japanese newspaper articles from Jan. 2003 to Dec. 2007. We used financial newspaper articles. The dataset consists of 6,984 publications, 546,230 open patents, and 16,364 newspaper articles. The categories assigned to publications are different from those of patent documents. More precisely, the number of categories assigned to publications is 155. The IPC(International Patent Classification) is assigned to each patent documents. We used second level categories consisting of 125 categories. Therefore, we manually selected

\(^3\)http://eigen.tuxfamily.org/index.php?title=Main_Page
\(^4\)www.jst.go.jp/EN/index.html
Table 1: Category correspondences.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Patent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer</td>
<td>Computation</td>
</tr>
<tr>
<td>Communication eng</td>
<td>Electronic communication</td>
</tr>
<tr>
<td>Oncology</td>
<td>Medical science</td>
</tr>
<tr>
<td>Electronics</td>
<td>Electronic eng</td>
</tr>
</tbody>
</table>

Table 2: Top 5 triples identified positive/negative.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>(ga geyougen koutyousha) company’s performance is good</td>
<td>(wo kahousiyousai shita) revise down sales</td>
</tr>
<tr>
<td>(ga mangokakatsuu atta) have demand granted in full</td>
<td>(ga shonshitu baizousita) debt increased</td>
</tr>
<tr>
<td>(wo seishinba shita) have capital investment</td>
<td>(wo futsai hokokusita) report a debt</td>
</tr>
<tr>
<td>(ga rieki baizoushita) profit increased</td>
<td>(ga sonsitu hettta) damage is decreased</td>
</tr>
<tr>
<td>(ga uruagedakar fuetto) the amount sold increased</td>
<td>(ga monai shinkokuda) The problem is so serious</td>
</tr>
</tbody>
</table>

a limited number of categories assigned to the publication that correspond to patent documents. Table 1 shows four categories used in the experiments.

We used plda\(^5\) to assign positive/negative label to the sentences in news reports. We estimated the number of topics (clusters) by searching in steps of 100 from 200 to 1,000. As a result, the minimum entropy value and the number of topics is 0.328, and 500, respectively. We used these in the experiments. Table 2 shows sample clusters regarded as positive and negative obtained by LDA. Each cluster shows the top 5 triples including verb/adjective that did not appear in the sentiment polarity dictionary. We observed that the extracted triples show positive opinion. This indicates that sentiment analysis contributes to classify news report into positive or negative. In contrast, a triplet such as (ga sonsitu hettta) “damage is decreased” which incorrectly classified into negative cluster is an obstacle to identify negative news reports in SVMs classification.

All news reports were parsed by the syntactic analyzer CaboCha(Kudo and Matsumoto, 2003), and 243,528 dependency triples are extracted. We used them as test data. We randomly selected another 600 news reports, and used them as a training data. As a result, the macro-averaged F-score concerning to positive and negative were 0.839, and 0.436, respectively. We added these news reports classified by SVMs to the original 600 training news reports, and used them as a training data to classify the total number of 3,000 test news reports. We obtained the total number of 2,053 positive news reports including positive training data, and used them to create temporal series.

We test our method by using closed and open data. In the experiment using closed data, we created temporal series by using all publications, open patents, and newspaper articles. These data include 585 companies in all. The difference time period between training and test data is one year and three years. When the time difference between training and test data is one year, we used the data collected from 2003 to 2006 years as a training data, and 2007 year as a test data. When the time difference is three years, we used the data collected from 2003 and 2004 years as a training, and the data collected from 2007 year as test data. We used five cross validation to evaluate the method. The estimated parameters are \(t=10\) (the number of eigenvalues), \(\alpha=0.5, \beta=0.3, \gamma=0.2\) in Eq. (2), \(\lambda=0.1\) in Eq. (4), and \(k=14\) of k-means\(^6\).

In the experiment using open data, we divided each of the three data into two, training and test data. Furthermore, we divided training data into two: training data to estimate parameters, and training data to construct prediction models by using these estimated parameters. The estimated parameters are \(t=10, \alpha=0.1, \beta=0.5, \gamma=0.4, \lambda=0.1\) and \(k=36\). We evaluated prediction performance by using three measures, i.e. mean-averaged absolute error(MAE), and mean-averaged relative error(MRE), and the coefficient of determination\(R^2\) (Yan et al., 2012) which are given by:

\[
MAE = \frac{1}{N} \sum_{t}^{N} | y_i(t) - f_i(t) | \tag{6}
\]

\[
MRE = \frac{1}{N} \sum_{t}^{N} \frac{| y_i(t) - f_i(t) |}{f_i(f)} \tag{7}
\]

\[
R^2 = 1 - \frac{\sum_{t}^{N} (y_i(t) - f_i(t))^2}{\sum_{t}^{N} (y_i(t) - y(t))^2} \tag{8}
\]

\(N\) shows the number of categories in the company we want to predict. \(\overline{y}(t)\) in Eq. (8) refers to the mean of the observed data at time \(t\), \(y_i(t)\) indicates each observed value and \(f_i(t)\) shows each predicted value. Eq. (8) shows that the value 1 indicates the best result, and the closer value to 1 shows better result. On

\(^5\)http://code.google.com/p/plda

\(^6\)For each category, we used the same value of the parameters.

287
the contrary, the smaller value of Eq. (6) and Eq. (7) indicates better result.
We recall that we extracted positive news reports from news articles, clustered temporal series, and estimated impact by citation network. To examine each method’s effectiveness, we compared our method with three baselines, i.e., the method not using news reports (Wo News), the method without SP clustering (Wo SP), and the method not using impact estimation by citation network (Wo IM). The results by both closed and open data are shown in Table 3.

We can see from Table 3 that the results obtained by our method were better than other methods in three evaluation measures. Moreover, out method was better than other methods in both of the time differences between training and test data, one and three years. We note that when we used open data, the difference between Ours and Wo SP was largest in all of the three evaluation measures. This shows that clustering temporal series by the SP clustering especially contributes to prediction. This is reasonable because the method without SP clustering predicts only one polynomial regression model, while there are variation of company’s trend on R&D. Figures 3 and 4 illustrates the result of company’s trend predicted correctly and incorrectly, respectively.

![Figure 3: An Example of correct prediction (Computer).](image)

![Figure 4: An Example of incorrect prediction (Electronics).](image)

X-axis in both Figures 3 and 4 indicates a time period of training data, 2003 and 2004 years, and a test data from 2007 year. Y-axis shows the specific company’s trend obtained by Eq. (4). A higher value of the company’s trend indicates more attractive in some business area. The business areas in Figures 3 and 4 is computer, and electronics, respectively. Doted line shows observed value, and a solid line refers to predicted value. We can see from Figure 3 that two lines are very close to each other in a prediction year, 2007. In contrast, Figure 4 shows that in year 2007, the electronics area in the company is more attractive while our prediction result shows that it becomes a slow curve incorrectly.

There are no existing work related to prediction of company’s future trend. Therefore, we compared our method to the most recent state of the art citation-based work related to high impact academic paper prediction proposed by Davletov et al (Davletov et al., 2014). More precisely, we applied Davletov’s method (Citation) to open patents and compared it with the result obtained by our method. Because in our data, only the open patents corpus has citation information. The results are shown in Table 4.

<table>
<thead>
<tr>
<th>Method</th>
<th>MRE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wo News</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wo SP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wo IM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4 shows that in both of the closed and open data, the results obtained by our method was better to those obtained by citation based method except for three years of time difference between training and test data in closed data by using MRE measure. When the creation time period of the test data differs three years from that of the training data, the overall performance of prediction decreased. However, the results obtained by our method were still better to those obtained by citation based method. This shows that prediction model obtained by combining three data is effective to predict company’s future trend.

4 CONCLUSION

We proposed a method for predicting company’s future trend on R&D. We used three types of data collections, scientific papers, open patents, and newspaper articles to estimate temporal change of trends on company’s business area. The results by using Japanese data show that the prediction model obtained by combining three data is effective to predict company’s future trends, especially the results show that SP clustering contributes overall performance.

There are a number of directions for future work. In the sentiment analysis, we should consider negative news articles to estimate a declining trend, and we should be able to obtain further advantages in efficacy by overcoming the lack of sufficient news reports by incorporating transfer learning approaches (Dai et al., 2007). Moreover, it is necessary to improve positive
news classification by smoothing terms which do not occur in the dictionary. We used Japanese newspaper articles to extract positive news reports in the experiments, while the method is applicable to other textual corpus. To evaluate the robustness of the methods, experimental evaluation by using other data such as Nikkei technology can be explored in future. Finally, comparison to other related work which make use of textual corpus (Joshi et al., 2010; Yagatama et al., 2011) will also be considered in the future.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their comments on the earlier version of this paper. The authors are also to thank Japan Science and Technology Agency for making available the data.

REFERENCES


