Amirkabir University of Technology
(Tehran Polytechnic)
Vol. 47, No. 1, Spring 2015, pp. 17- 22

Amirkabir International Journal of Science& Research
(Electrical & Electronics Engineering)
(AIJ-EEE)

Automatic Discovery of Technology Networks for Industrial-Scale R&D IT Projects via Data Mining

S. Azimi[1], H. Veisi[2*] and R. Rahmani[2]

1- Faculty Member, Department of New Sciences & Technology, Tehran University, Tehran, Iran
2- Assistant Professor, Department of New Sciences & Technology, Tehran University, Tehran, Iran

ABSTRACT

Industrial-Scale R&D IT Projects depend on many sub-technologies which need to be understood and have their risks analysed before the project can begin for their success. When planning such an industrial-scale project, the list of technologies and the associations of these technologies with each other is often complex and form a network. Discovery of this network of technologies is time consuming for a human to perform, due to the large number of technologies and due to the fact that the technologies are constantly changing. In this paper, a method is provided for the automatic discovery of the network of associations of Industrial IT technologies as a networked graph, using data mining and web-mining algorithms. The proposed process is an approach to form a dynamic weighted graph of technologies. A numeric value is calculated as similarity between technologies. A combination of data mining and web mining techniques have been used to achieve the results. The main objective is to invent a computerized reproducible method so that by the help of it, technological relation can be extracted and updated constantly. This method consists of six phases, of which four phases are performed automatically by novel algorithms introduced in this paper. The analysis of more than 8 million terms suggests that the proposed method provides acceptable results. This paper also provided recommendations to improve the suggested method**.**

KEYWORDS

Technology Mining, Technology Graph, Data Mining, Text Mining, Similarity Algorithms, Web Robots.

* Corresponding Author, Email: h.veisi@ut.ac.ir

1-INTRODUCTION

One in six IT projects run out of control with an average run topping %200, and a schedule overrun of almost 70% [1].

In Industrial-Scale R&D IT Projects, there are many management and technical issues in projects including prioritizing research sub-projects, determining the effectiveness of a propellant action, highlighting the co-operation of the research project and choosing the right organizational partnerships for economic efficiency.

One of the requirements for definition and management of technological research projects is to detect the technologies involved in a project accurately. Technology area is one of the main areas affecting the project's risk [2]. Finalizing this step improperly leads to serious challenges in project estimation and its feasibility. As a result, the possibility of achieving an acceptable outcome reduced.

National Health Service project in UK had more than 4 years delay and over 5 time over cost ($4.6B to $24B), the project faced with some technical faults [3].

When the technological scope of a research project is not specified, the risk enhanced and managing would be a very difficult task. In such cases, there are two situations: in the first case, project budget and duration could not be increased so the goals would not be achievable. In the second case, when it is vital to attain the project objectives (this is more important than project budget and duration) permanently new issues are raised during the project and new branches are created so the project managing is a fully developing chaos. If the associations between the technologies involved in the project are diagnosed properly, the technological scope of the project will be determined.

So far, different methods for analysing technology have been defined and sometimes implemented [4]. For example, methods such as technological intelligence, forecasting, roadmap, estimation and technology foresight. In [3], a combination of different methods of technical analysis is presented. It should be noted that this method, like many others, in contrast to the technological dynamism, takes so much time. In this way, when it attained a result, it has become outdated before. In order to solve such problems, a new model is proposed to analyse the technology [5]. It involves several processes; the first step is to analyse the technologies by accessing an informative database. Then, an automated decision-making process is involved. In this method, there is little emphasis on data collection for the database. This can be considered a challenge for the analysis of the updated technology.

In this paper, we provide a method for finding the relationship between technologies. In this method, data are collected through an innovative techniques based on the web mining. Then after some steps of preparing, the similarity algorithm was executed on data. Professor Belver Griffith in one of the first innovative methods experts who has used relation between citations in papers [6]. Using bibliometric for the analysis of scientific-research output is another method. Bibliometric utilizes data which are called as “published data”. In this way, different fragments of the article distinct to physical parts and each one will be analysed separately [7]. Goals in [8] are considered comparable to the goals of our study; there have been attempts to evaluate technologies and gaining a potential technological development for each. This leads to a procedural titled "Technology Opportunities Analysis". In this method, the analyses of information in “INSPECT” database besides data mining and data monitoring are utilized. Finally, the statistical measures are employed, too.

The aim of the present study is to establish an iterative process to extract the relationship between technologies. Recording technologies related to a Specific technology and then determine the weights of these relations are expected, too. On the other hand, due to the dynamic nature of the various elements in the technology space, we have tried to define all phases recursively. As the/a result, the process of designing and implementation of the iterative algorithm could be repeated at any time so the results would be updated.

According to this method, the work is done in six phases defined as follows:

1. Initial data collection (web robot design and implementation, besides execution of collecting steps)
2. Data refinement
3. Extraction of repeated phrases
4. Extraction of combined repeated phrases
5. Similarity calculation
6. Results analysis

In the second section of this article, along with their algorithms, the five implementation phases are described. At the end of the second section, in the sixth phase, some of outputs are represented. In the third section, after examining the output of operational phases, issues and possible improvements are revealed.

2-FIND SIMILARITY BETWEEN TECHNOLOGIES

2-1- PHASE 1:

INITIAL DATA COLLECTION

The same as the results of similar studies in Ireland University in [9] by now, extracting the relationship between technologies has not been an automatic process yet. In our case, human intelligence is more crucial in two phases: setting inputs and analysing outputs. In the first phase, setting the data as inputs and refinement of them are performed. The main objective of this phase is to create a repeatable process to collect raw data. This phase is conducted in three parts:

- Select a source to extract a list of technologies
- Review and refine the list of technologies
- Build a web robot

In the first part, several sources which have categorized IT field technologies have been studied [10] [11] [12]. Among the possible choices, Webopedia website [13] as a complete reference is selected and a database of all the terms associated with the technology is extracted from it. In implementation, Postger SQL as database tool and Java programming language are used.

The initial database contains 10,424 tech titles. In the second part of the first phase, the resulting list was refined. So phrases like "deploy" or "scalable" are excluded from the list. Then duplicate terms are also eliminated. At the end, 6526 terms are obtained for the next step.

In the third part, a resource for data collection should be selected. Based on the scientific validity, extension and being updated, IEEE [14] is selected. Then a web robot is designed and implemented to collect the relevant data. The robot searches each phrase in the IEEE website and save the abstracts of founded papers in a database. Since the search engine of IEEE website shows papers according to the relation to the search term, to limit the obtained data, the first 500 search results are chosen. Before, some other methods to search in IEEE were tried; for example, searching in keywords given by authors. This approach does not seem appropriate. According to limitation of keywords number (5 to 6 words) in comparison with our method

For ensuring the search results to be updated, time period has been set between 2012 and 2014. Regarding to the above configuration, robot web application has been designed, tested and then run. Finally, 836,498 abstracts were identified according to 6,526 chosen terms.

2-2- PHASE 2:

DATA REFINEMENT

One of the requirements of data mining process is eliminating stop words or extra data. Likewise, according to the data collected by web robots, noise in collected data was clear. Thus, by using various resources, a database for standards stop-words were prepared. According to the essence of information technology, input data includes phrases such as "IT" or "VIA" (Versatile Interface Adapter) misdiagnosis with regular expressions of "it" or "via". Therefore, it would be necessary to simply remove "it" with a lowercase letter. But if "it" exists in the beginning of a sentence and showing up "It" still we need to remove it. That is why the phrase "it" and "It" both add to stop words database.

The phrases that represent special characters, like &gt instead of "<" symbol, have been added to the database. Finally, by adding similar noise seen during several executing and outputs checking, stop-words database completed. At the ends, its records reach 1353.

2-3- PHASE 3:

REPEATED PHRASES EXTRACTION

In data mining, to extract the rules, recurring patterns are always tried to be identified. Then, according to them, rules and environment procedures are discovered [15]. In the present method, by defining several characteristics, a benchmark to compare similarities has been utilized.
The first feature is Frequent Item Set (FIS). For this purpose, all relevant abstracts related to a term had been reviewed. Then its FIS besides the number of repetitions were recorded in a separate table. In this case, for every 6,526 technologies, a list of their FIS (average 1235 words) was extracted. This just included FIS repeated at least twice. Finally 8,062,897 FIS were recorded in the database. To lower the calculation, the min support was supposed to be "10". In this way, for each technology term, about 220 FIS with 10 or more repetitions have been recorded. In this phase, for accelerating counting applications, the Hash map of Java is utilized.

2-4- PHASE 4:

EXTRACTING COMBINATORIAL FIS

As the main target of this project had been established based on finding the association between technologies; Mining the combinatorial FIS has been assumed as the most practical and useful way. In this case, FIS including two or three words have more score than FIS including single worked. In this respect, some data mining algorithms and tests have been considered and implemented. The results have clarified that algorithms like FPgrows [16] and Aprioriall [17] can be regarded as beneficial to the market transactions. In market data models, the numbers of the fields are limited but there is numerous of records. However, unlike the market data models, in this circumstance, there are lots of properties (words of abstracts) and a few records (number of abstracts related to a specific title which are at most 500 abstracts). Accordingly, it seems that some specific text-mining algorithm should be utilized. In this regard, several methods have been anatomized to propel the project. Nevertheless, consequently, one innovative method was designed which has been initiated based on the method [18]. In this method, instead of making the collection of the combination of FIS which finally leads to numerous records, the invented algorithm has been exploited to concentrate on FIS of specific title. Firstly, titles included more than 9 frequent thresholds selected (3795 titles) and then in the second step; the set of FIS's of each title has been chosen. Meanwhile, in the third step, FIS's were searched in all obtained abstracts of title. When for the first time each FIS was found in an abstract, two tables were formed. The first one contains the FIS plus the next word in the abstract and in the second table, the FIS plus two next words were inserted as a nominated 3-segment FIS. By continuing the algorithm before inserting new items, the existence of them was checked and in the case of their existence, instead of insertion, just one unit was added to the counter of repetition.

To increase the speed of algorithm, initially, the first and last place of FIS in abstract have been obtained. If the first position was zero, it would mean that the FIS has not been found in abstract. If the position of first and last was the same, it would mean that there is only one occurrence of FIS in abstract. Beyond these two supposed states,

only the place between the first and last one has been searched.

If a three-word phrase has been frequent, undoubtedly, the phrase of two first words and two last words will be frequent, too. Nonetheless, this frequency should not be counted as two part frequent words. In a section of algorithm, this issue is mentioned and defected.

The algorithm has been performed over 1,438,668 words of FIS and more than eight hundred thousands of linked abstracts by a Core i5 computer equipped to 8 GB Ram. This task takes around less than 2 hours and leads to gain the following outputs:

- 118,037 two part frequent phrases
- 17,849 three part frequent phrases
- The pseudo code of algorithm is depicted in Fig-1

```
For every_terms
  A_abstracts=term.all_abstracts
  A_fIS=term.all_FIS
    For each i in A_abstracts.next
      A_term_words= Tokenizer(A_Abstaracts)
        For each j in A_FIS
          A=first_indexof(A_FIS(j)) in A_term_words
          B=last_indexof(A_FIS(j)) in A_term_words
            For (k=A to B)
              If A_FIS(j) = A_term_words(k)
              Put (A_term_words(k) +A_term_word(k+1)) in
                - twin list or if exist ++counter
              Put (A_term_words(k)+A_term_word(k+1))+A_
                - term_word(k+2)) in triple list or if exist ++counter
  A_term_words.clear
```

Fig. 1. Pseudo code of algorithm

2-5- PHASE 5:

SIMILARITY CALCULATION

At this stage, three features have been prepared to calculate the similarity. These features are: frequent single words, frequent two-word phrases and frequent three-word phrases. For each of the tech titles, some of these features were stored in the database.

The similarity between two tech titles has been considered as a reciprocal relationship. It means that the similarity between A and B technology is equal to the similarity between B and A technology. As a result, the matrix of similarities is a triangle matrix. In this respect, Fig 2 depicts a model of this matrix which all 3775 titles of tech have been crossed with each other.

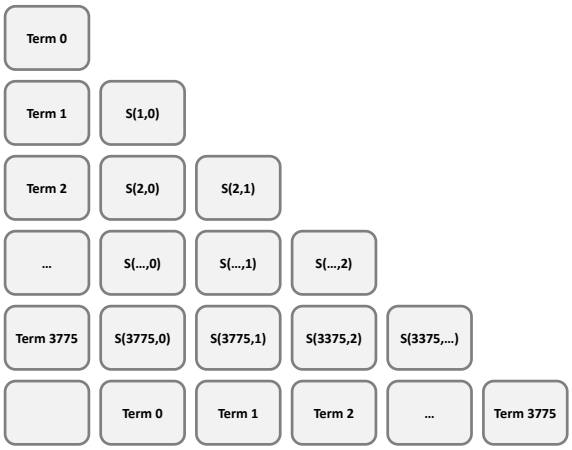

Fig. 2. Matrix of similarities

In such corresponded model by considering the n.(n-1)/2 formula, it was required to calculate the value of 7,123,425 similarities. Calculating each similarity takes around 0.96 second thereby, the whole calculation takes around 80 days. As it has been aforementioned in section 2-4, two and three part phrases have more score than single word phrases in calculating the similarity. Therefore, after examining several coefficients and the output results, formula (1) was designated for the calculation.

$$similariy = \frac{C_{si} \times \left(\frac{com_{si}}{T_{si}}\right) + C_{tw} \times \left(\frac{com_{tw}}{T_{tw}}\right) + C_{tr} \times \left(\frac{com_{tr}}{T_{tr}}\right)}{c_{si} + c_{tw} + c_{tr}} \quad (1)$$

Where:

C_{si}	Coefficient of single words
C_{tw}	Coefficient of two words phrases
C_{tr}	Coefficient of three words phrases
com_{si}	Number of common single words FIS
com_{tw}	Number of common two words phrases FIS
com_{tr}	Number of common three words phrases FIS
T_{si}	Total FIS of single words
T_{tw}	Total FIS of two words phrases
T_{tr}	Total FIS of three words phrases

This formula first normalizes all of the achieved features. Hence there is this potentiality that for example two phrases which each one might contain 1000 FIS and only 50 of them will be the same. Alternatively, two other phrases each one may have only 100 FIS but 25 of them will be the same. It is obvious that the similarity of two pairs in the second example is more and so, it is compulsory to normalize the number of similar phrases.

In this formula; after normalization of each feature, the coefficient of each feature will be applied and finally, for each feature, the average of resulting values will be calculated.

EXPLANATION OF THE ALGORITHM

The designed algorithm for this phase consists of four loops in which the function of similarity is executed in

the fourth and internal loop (fig 3). To simplify the code showing, reading all FIS of a tech title is briefed as "For each terms(i). FIS". This one line pseudo code is encapsulated in three functions. Reading single word FIS and storing them in an array, reading two words FIS and creating separate array for them, and finally the same thing has been implemented for the three word FIS's. After this initialization, the function of similarity will be performed on each pair of tech titles.

```
For i=1 ; i <= terms.count
  For j=terms(i+1) ; j <= terms.count
    For each terms(i).FIS
      For each tersms(j).FIS
        Calculate Similarity(i,j)
```

Fig. 3. Similarity calculation loops

- According to algorithm which was described in phase four, "Finding first and last position" method is used to increase the speed of execution.
- Outputs are stored in separated Hash maps and at the end of each loop, stored in database in a batch transaction.

2-6- RESULTS ANALYSIS

After the conclusion of the phase 5, similarities for each tech titles have has been calculated and for each one of tech titles, 3774 records have been inserted. Table 1 contains similar technologies with descending similarity values. To have a better view, only 10 top similar technologies have been illuminated. This table also demonstrates only 3 instances.

TABLE 1. SORTED SAMPLE OF SIMILARITY CALCULATION

Tech Title	Title of related tech	Similarity
data center	data server	0.386552
data center	data center management	0.340339
data center	link server	0.235896
data center	VMS	0.210074
data center	Linked Data	0.204438
data center	virtual machine server	0.203387
data center	cloud migration	0.201857
data center	network access server	0.195424
data center	routing switch	0.192592
data center	virtual server	0.189258
IaaS	application virtualization	0.805381
IaaS	VMS	0.800813
IaaS	hosting services	0.612836
IaaS	CAMP - Cloud Application Management for Platforms	0.593467
IaaS	cloud migration	0.576292
IaaS	virtual server	0.573612
IaaS	server virtualization	0.572393
IaaS	virtual machine	0.5715
IaaS	hosting server	0.567989
IaaS	virtual machine server	0.555299
mobile broadband	fixed wireless	0.749599
mobile broadband	802.16	0.740545
mobile broadband	MAC Layer	0.729693
mobile broadband	IEEE 802 standards	0.716959
mobile broadband	IEEE	0.661189
mobile broadband	MIB	0.55155
mobile broadband	WiMAX	0.495156
mobile broadband	broadband	0.492771
mobile broadband	wireless	0.492185
mobile broadband	LAN - local-area network	0.459209

The results illustrate that the famous idioms "input rubbish, get rubbish" are totally working here. Incorrect tech title as input is the main reason to have an irrelevant output. As it has been shown in table 1, it should be clarified that the term "IEEE" is a tech title or not. Irrefutably; there are different answers to this question based on miscellaneous viewpoints. If the right view from the specific end user had been selected, better results would achieve. As it has been illustrated in table 1, the term "IEEE" has been designated the fifth rank to itself in comparison with the similar title "mobile broadband". Here "IEEE" is a set of standards and protocols for "mobile broadband" and to work on that, these standards and protocols should be considered. In this case, such perception seems implausible of artificial intelligence. However, such reminder may be helpful for a Tech Development manager of a company.

Another key point is the usage of abbreviations as tech titles. Results which show similar titles with HSP were unlike (table 2). The first investigations show that HSP is the abbreviation for several phrases such "hosting Service Provider", "Host signal processing" and "Hot Soup Processor". Thereby, there were no precise and alike outputs for this title. Similarly, the similarity values which obtained were not more than 0.133094.

TABLE 2. TOP FIVE SIMILAR TITLES WITH HSP

Tech Title	Title of related tech	Similarity
HSP	RAT	0.133094
HSP	Windows 7	0.131894
HSP	Windows 7 HomeGroup	0.131894
HSP	TC	0.130695
HSP	eBay Compatible Application	0.129496

When full title instead of abbreviation was used, the output was more precise and the similarity values were more than the number in table 2. It can be anatomized in table 3.

TABLE 3. TOP FIVE SIMILAR TITLES WITH HOSTING SERVICE PROVIDER

Tech Title	Title of related tech	Similarity
Hosting Services Provider	Virtual IT Service Provider	0.828487
Hosting Services Provider	virtual machine server	0.805882
Hosting Services Provider	hosting services	0.699071
Hosting Services Provider	hosting server	0.630408
Hosting Services Provider	CAMP - Cloud Application Management for Platforms	0.628806

3- CONCLUSIONS

The most effort to present this method was on creating an automated and complete cycle of finding similarity between technologies. Automation of this cycle makes it repeatable. After conclusion of operational phases, several issues and improvements were revealed in each phase. These issues and improvements start from the preliminary steps link "tools selection" and "basic methods" up to final stage "results analysis". A number of improvements are stated as follows:

- **Technology terms (primary inputs):** A list of technology titles have been selected as primary inputs. These lists can be improved greatly.

- **Improving the list of technologies:** Data mining on similarity results shows that the data which generated among the similarity calculation phase could be used to recognize the new technology titles. A suitable method in this case is now under production.

- **Lack of measures to estimate the accuracy of the method problem:** To understand the validity of the proposed algorithms, the output method has only been studied from the perspective of comparative and intuitive. It is necessary to measure the quality of the developed method for the quantitative determination.

- **Weights assignment (coefficients):** by definition of appropriate measures, along with the production and study of the outputs, intuitive method of determining the coefficients can be improved.

- **The issue of calculation time:** The proposed algorithm is very time consuming. Accurate determination of effective elements in the calculation, data summarization and using parallel processing may partially solve the problem.

The proposed idea and its implementation in this paper are the new methods to analyze the technologies. In this way, an innovative algorithm is implemented to extract similarity and associations between information. It does this by analyzing data from scientific papers on the IEEE website. Finally, by applying several improvements to the algorithm and their implementation, communication and similarity results are presented. It is important to note this algorithm is able to be executed at any time on the website. As a consequence, the relevant outcome could be obtained updated.

REFERENCES

[1] Flyvbjerg, Bent, and Alexander Budzier. "Why your IT project may be riskier than you think." Harvard Business Review 89.9 (2011): 601-603.

[2] Yacov Y. Haimes, Risk Modeling, Assessment, and Management, 2005, Wiley Books, 615-629

[3] https://en.wikipedia.org/wiki/NHS_Connecting_for_Health, viewed at 2014

[4] Alan L. Porter, "Technology futures analysis: Toward integration of the field and new methods", Technological Forecasting & Social Change 71 (2004) 287–303

[5] Alan L. Porter, "Quick Technology Intelligence Processes", eu-us seminar: new technology foresight, forecasting & assessment methods-seville 13-14 may 2004

[6] Belver c. Griffith –"Mapping the Scientific Literature." NATO Advances Study Institutes Series Volume 10, 1975, pp 457-481

[7] Melkérs, J., bibliometrics as a Tool for Analysis of R&D impact, in Evaluating R&D impacts: Methods and Pracüce, B. Bozeman and J. Melkers, eds., Kluwer, Boston, 1993, pp. 43-61

[8] Donghua Zhu, Alan Porter, Scott Cunningham, Judith Carlisie, Anustup Nayak. "A process for mining science & technology documents databases", illustrated for the case of "knowledge discovery and data mining". Ci. Inf. vol.28 n.1 Brasilia Jan. 1999

[9] Behrang QasemiZadeh, "Towards Technology Structure Mining from Scientic Literature"

[10] Information Technology Encolopedia, http://whatis.techtarget.com/glossaries, Viewed at 2014

[11] The Tech Terms Computer Dictionary, http://www.techterms.com/, Viewed at 2014

[12] Category:Information technology, http://en.wikipedia.org/wiki/Category:Information_technology, Viewed at 2014

[13] Tech dictionary for IT professionals and educators, www.webopedia.com, Viewed at 2014

[14] Advancing Technology for Humanity, http://www.ieee.org/, Viewed at 2014

[15] Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." ACM SIGMOD Record. Vol. 29. No. 2. ACM, 2000.

[16] Agrawal, Rakesh, and Ramakrishnan Srikant. "Mining sequential patterns." Data Engineering, 1995. Proceedings of the Eleventh International Conference on. IEEE, 1995.

[17] T. P. Martin and M. Azmi-Murad, An Incremental Algorithm to find Asymmetric Word Similarities for Fuzzy Text Mining, Soft Computing as Transdisciplinary Science and Technology Advances in Soft Computing Volume 29, 2005, pp 838-8

[18] Zhao, Qiankun, and Sourav S. Bhowmick. "Association rule mining: A survey." Nanyang Technological University, Singapore (2003).