Information Retrieval and Visualization for Searching Scientific articles and Patents

Lipika Dey, Hemant Gupta, Kunal Ranjan

Innovation Labs
Tata Consultancy Services
Delhi, India
lipika.dey@tcs.com, gupta.hemant@tcs.com

Abstract. Given the rapidly changing face of technology, keeping up with the trends and identifying potential areas to be explored for research or commercialization is a challenging task. Decision makers, research analysts, scholars, research directors all make use of digital collections, use of which is facilitated by search applications developed on top of them. However, search is a human-driven activity and the result of such analysis is largely dependent on the initial inputs that are provided by the expert. Besides, aggregating and assimilating all the information returned by a search engine is no less daunting. In this paper, we propose intelligent methods for presenting search results to help information assimilation. We also present methods for analyzing large collections of documents in an automated way to generate insights that can prove to be useful for analysts. Starting from time-stamped collections of research publications and patent documents, we present several Information retrieval (IR) techniques that can successfully extract and present insights about emerging, popular and receding trends in research along with their current levels of commercialization. We present results of experiments based on research abstracts made available by digital libraries and US patent office.

Keywords: Topic Extraction, Information Retrieval, Commercialization score

1 Introduction

As an off-shoot of the popularity and accessibility of world-wide web there has been a phenomenal rise in the number of research articles, conference proceedings, archived research results, patent filings and grants and several other technical publications which are available online. Though this collection is extremely useful for academic and industrial researchers, searching for any information results in a huge list of documents. Assimilating information from this huge list is a formidable task. While researchers query research collections to understand evolution of an area or topic, state of the art etc. business users may be querying the same collection to remain abreast of the latest in research and look for possible ideas of commercialization to retain competitive edge. In this paper, we present our work towards developing a search and

pp. 265–282; rec. 2015-01-25; acc. 2015-02-25

analytics system that can aid in retrieving relevant information and insight generation from an integrated collection of research publications and patent applications.

Advances in search and information retrieval technology ensure that large volumes of text can be searched efficiently when appropriate queries can be formulated. The focus of text mining research on the other hand has been towards insight generation from large collections of technical documents. Identifying new trends in research topics is a popular area of research. Identifying and exploring relations among research communities is also a popular area. Visualization of information is also a deeply researched area. Some authors have also studied topic evolution over patents and research publications.

However presently there is no search and analytics system that goes beyond listing of articles for an integrated collection of scientific publications and patent documents. Though a listing provides numerical assessment about the potential presence of articles, it does not allow users to easily perceive (a). Content-based relationship among different research areas either at theoretical or application level or (b). the true extent of commercialization of an area or topic. The utility of such a system can be manifold. It can help researchers understand the applicability of research topics. For strategists and decision makers, it would be of help to find yet untapped areas of research and potential areas of new application developments.

The unique aspects of the present paper are as follows:

1. A novel method is presented to identify topic evolution using topically significant phrases, where topics are extracted from time-stamped collections using standard Latent Dirichlet Allocation (LDA). The topical phrases are also used to present a graphical representation of how the underlying topics have evolved or morphed over the years. We have proposed new topic-similarity measures based on Information retrieval (IR) principles that take into account relevance of a document with respect to a topic, rather than word-based measures.
2. The paper proposes new measures to compute the extent of commercialization of a research topic with respect to a patent database. We term this as commercialization score of a research topic. While we have conducted experiments and presented results from the US patent database for the sets of applied and granted US patents over the years 2005 to 2014, the measures are generic and can be used in conjunction to any such database.
3. We present a method for analyzing commercialization scores and commercialization trends to generate insights about further prospects of a topic or an area.

The rest of the paper is organized as follows. Section 2 provides a review of related work. Section 3 discusses how topic similarities are computed to generate a topic evolution graph. Section 4 presents the proposed methods to compute commercialization score and commercialization trends. Section 5 presents some results obtained over a publicly available data set. Finally section 6 concludes with future work.

2 Review of Related Work

A large number of research communities are actively engaged in analyzing scientific articles and patent applications. An interactive prototype system named Action Science Explorer (ASE) was presented in [1], to help researchers with reference management, analyzing topical and citation statistics, text extraction and natural language summarization for single and multiple documents. It supported network visualizations to see citation patterns and identify author clusters. ArnetMiner was proposed in [2]. This paper proposed a unified tagging approach using Conditional Random Fields to generate profile tags for researchers based on publication data extracted from the web. It also proposed a unified topic model called Author-Conference-Topic (ACT) to simultaneously model different types of information in the academic network. Rexplore [4] supports graph-based exploration to understand bibliographic data, research topics and trends. It exploits the Klink algorithm [3] which identifies relations across different research areas using semantically annotated data. [5] proposed several metrics of influence, coverage, and connectivity for scientific literature which can be used to create structured summaries of information, called metro maps. Metro maps are targeted at capturing the developments in a field. An iterative topic evolution learning framework was proposed in [17] based on an inheritance topic model that leveraged citations among documents to analyze topic evolution in an explicit way.

Several groups have also tried to capture researcher communities and group dynamics [6-8] from content and not just from citation. [6] used a key-word based approach to identify topics. In [7] which is an extension of [4], authors proposed the notion of diachronic topic based on communities of people who work on semantically related topics at the same time. It was used to detect events that denote topic shifts within a research community; the appearance and fading of a community; splitting, merging and spawning of new com-munities etc. [8] presents a detailed study on the factors that affect research collaboration among individuals and organizations.

[9] presents a comprehensive literature review on research around analysis of patents. A topic-driven patent analysis and mining system was presented in [10] which studied the evolution of patent network composed of companies, inventors, and technical content using dynamic probabilistic model. It also proposed analytics tools for IP and R&D strategy planning, including a heterogeneous network co-ranking method, a topic-level competitor evolution analysis algorithm, and a method to summarize the search results. [11] proposed an analytical technique called patent trend change mining (PTCM) to capture changes in patent trends. This work, based on association rule-mining was aimed at generating competitive intelligence to help managers develop appropriate business strategies based on their findings. [12] presented a patent analysis system called TechPerceptor which used Natural Language Processing techniques to generate patent maps and patent net-works based on semantic analysis of patents. The system can be used to observe technological hotspots and spot patent vacuums. [13] proposed the use of text mining techniques to develop a Technology Tree(Tech Tree) that can compute similarity scores between patents.

None of the existing systems perform joint analysis of publications and patents using the content of both publications and patent applications. Most importantly none of the systems provide insights about the extent and diversity of research topics and their commercialization to help technology planners.

3 Topic Evolution and Diversification

Latent Dirichlet Allocation [14] is an unsupervised latent variable model that employs Bayesian inference to identify semantic clusters of words in document collections that resemble topics. LDA assumes a range of possible distributions of words with the constraint that they are drawn from Dirichlet distributions. This enables it to learn latent topic models in an un-supervised way ensuring that the topic models are maximally relevant to the underlying data collection. For the proposed work, the LDA model was first applied on yearly collections of publications, which yields topic distribution for each document. Each topic comprises bag of words along with probabilities of each word being generated by that topic.

Word-based representation of topics is useful, but not easy to understand. Instead, the present system adopts phrase-based representation of topic that was proposed in [16]. For each topic, its representative phrases are chosen from among frequently occurring three-grams and two-grams in documents that have a high probability of that topic. Since each document has a probability of each topic being present in it, [16] presented equations to compute the probability of a phrase belonging to a topic based on the occurrence frequency of phrases within documents that contained the topic with a probability greater than a pre-specified threshold. The maximally weighted phrase is used to name the topic. Phrases in the current context refer to N-grams that are faster to compute than natural-language phrases and are also resistant to noise like incorrect grammar or incorrect formatting. N-grams also preserve spatial relationship of words thereby making them closer in appearance to natural language phrases though obtained at much lower computational cost. The frequent n-grams selected to represent a topic are termed as topical phrases. Figure 1 shows phrase based representation of topics that contained the phrase “association rule mining” over the years 2006 to 2009.

Figure 1 shows that a research topic does not remain static over the years. Topics grow, evolve and diversify. A topic’s growth can be tracked by watching the trends in number of publications that continue to cover the topic. Topic evolution can be tracked by watching the changing content. This cannot be tracked using simple word-based representation of topics since the words are difficult to interpret without their context. For example, the word “information” can make many topics look similar, though in reality the topics “Information Retrieval”, “Information Security” and “Management Information Systems” are quite different. Also, new words or phrases emerge and become frequent while old ones phase out. It is therefore proposed that co-occurrences of phrases can better capture continuation and evolution of topics.

Topic diversification captures inter-mixing of topics or adoption of a topic into another topic etc. Figure 2 presents year-wise view of frequently co-occurring N-grams for the query “association rule mining”. It may be noted that the context of “association rule mining” is different from its topical representation shown in figure 2. In fact the 2010 collection did not yield a topic named “association rule mining” though the phrase occurred in the context of “genetic programming” and “traffic prediction”. This obviously indicates that areas like “intrusion detection” or “web traffic prediction” had started adopting association rule mining techniques from 2009 onwards.

We now present a new method to capture topic similarity and then go on to show how this can be used to capture topic evolution and diversity.
Let T_i and T_j represent two different topics of the same year or different years. The topical similarity between T_i and $T_{j,}$ denoted by $\sigma(T_i, T_j)$, is computed in terms of their topical phrases as follows.

Let S_i and S_j be the sets of top n topical phrases associated to T_i and T_j respectively. Let p_i and p_j represent two phrases where $p_i \in S_i$ *and* $p_j \in S_j$.

Let D_i *and* D_j denote the collections of documents that contain p_i and p_j respectively. D_i *and* D_j may be identical, overlapping or completely disjoint. The degree of overlap of these two sets captures the *neighborhood similarity* of p_i and p_j, denoted by $\eta(p_i, p_j)$ and is computed as follows:

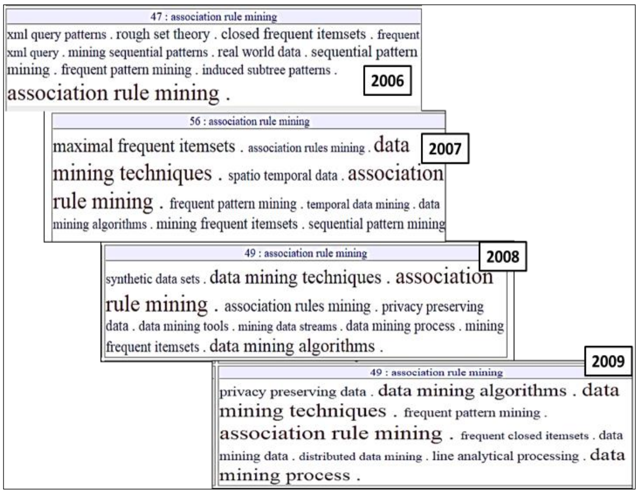
$$\eta(p_i, p_j) = \frac{|D_i \cap D_j|}{|D_i \cup D_j|} \tag{2}$$

For each phrase $p_i \in S_{i,}$ let $\alpha_j \epsilon S_j$ be the phrase with maximum value for $\eta(p_i, \alpha_j)$ i.e. $\eta(p_i, \alpha_j) \geq \eta(p_i, p_j)\ \forall p_j \in S_j$. In other words, the phrase p_i of topic T_i co-occurs maximally with α_j of T_j. Similarly, for each phrase $p_j \in S_j$ let $\beta_i \epsilon S_i$ be the phrase with maximum value for $\eta(\beta_i, p_j)$ i.e. $\eta(\beta_i, p_j) \geq \eta(\beta_i, p_j)\ \forall \beta_i \in S_i$.

It is obvious that the neighborhood similarities for a pair of phrases are not symmetric in nature. The similarity between a pair of topics is computed as the average neighborhood similarity between all pairs of topical phrases for pair.

$$\sigma(T_i, T_j) = \frac{1}{2n}\left(\sum_{i=1}^{n} \eta(p_i, \alpha_j) + \sum_{j=1}^{n} \eta(\beta_i, p_j)\right) \tag{3}$$

It may be noted that unlike most similarity measures that are computed on the basis of shared words or terms, $\sigma(T_i, T_j)$ computes similarity of topics in terms of shared documents in which representative terms of T_i and T_j co-occur.



47 : association rule mining

xml query patterns . rough set theory . closed frequent itemsets . frequent xml query . mining sequential patterns . real world data . sequential pattern mining . frequent pattern mining . induced subtree patterns . association rule mining .

2006

56 : association rule mining

maximal frequent itemsets . association rules mining . data mining techniques . spatio temporal data . association rule mining . frequent pattern mining . temporal data mining . data mining algorithms . mining frequent itemsets . sequential pattern mining

2007

49 : association rule mining

synthetic data sets . data mining techniques . association rule mining . association rules mining . privacy preserving data . data mining tools . mining data streams . data mining process . mining frequent itemsets . data mining algorithms .

2008

49 : association rule mining

privacy preserving data . data mining algorithms . data mining techniques . frequent pattern mining . association rule mining . frequent closed itemsets . data mining data . distributed data mining . line analytical processing . data mining process .

2009

Fig. 1. Topic - "Association Rule Mining" phrases through years 2006 – 2009

Topical evolution is captured through intra-year and inter-year topic similarity matrices. An intra-year topic similarity matrix captures pair-wise topic similarities for topics belonging to the same year. An inter-year similarity matrix captures pair-wise similarity for topics of consecutive years. Thus for a time-stamped collection containing articles published over N consecutive years, we obtain N intra-year similarity matrices and N-1 inter-year similarity matrices, each of k^2 dimension, where k is the number of topics per year.

The similarity-matrices constructed as above can be considered as adjacency-matrix representation for a multi-layered labeled and weighted graph G in which nodes represent topics. Each layer contains nodes representing topics of the same year. Nodes within a single layer are connected by weighted, undirected edges where the weight of an edge is equal to the similarity of the topics connected by it. Absence of an edge indicates no similarity. A pair of nodes from two different layers is connected by a weighted edge if the layers denote consecutive years. The weight is again equal to the similarity of the two topics it connects. A node in this graph is denoted by T_i^m where *i* is a topic index and *m* is a year-index.

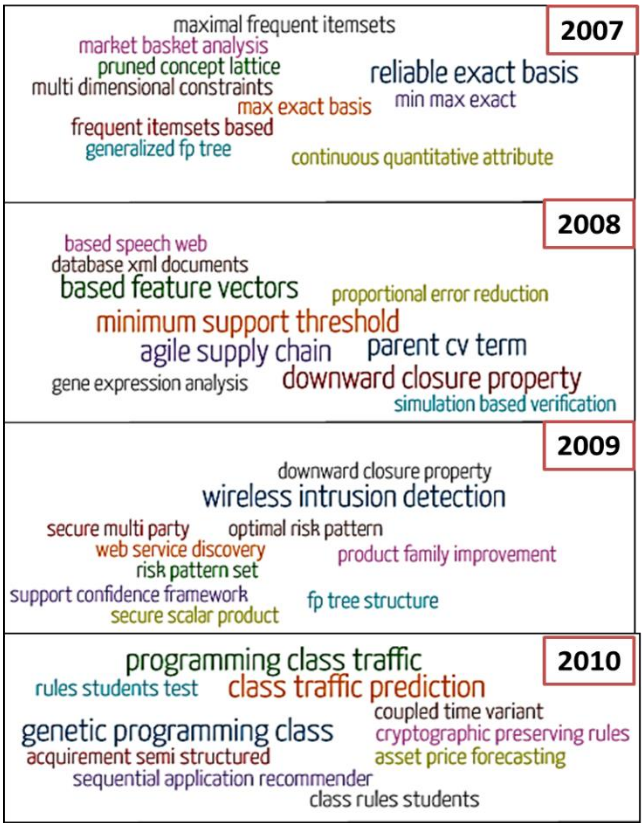


Fig. 2. Evolution of "Association Rule Mining" through contextual collection of topical phrases

It is proposed that topic evolution and diversification can be obtained as strongly connected components of the above graph. The algorithm proposed below finds strongly connected components within the layered graph. It uses two parameters ε and κ, which are defined below.

Definition 1: ε is defined as the *similarity_threshold* of the topic evolution graph. Two topics T_i, T_j are considered to be ε-related if and only if $\sigma(T_i, T_j) > \varepsilon$. ε-related does not imply that one topic has evolved from another topic. It is the minimum requirement for evolution.

Definition 2: κ is defined as the *connectivity_threshold* for topic evolution. The value of κ lies between 0 and 1. A set of λ nodes are said to be κ_connected to each other, provided each of them is ε-related to at least κλ number of nodes from this set. When κ is equal to 1, the set of nodes are fully connected to each other.

We now explain the algorithm to find κ_connected components of ε-related topic-similarity matrix.

4. Input ε and κ. Initialize C to NULL. C will finally contain a set of independent components, where each component will denote a set of connected topics.

5. Let B = (D, E) be a sub-graph of G which is constructed as follows. E contains only those edges of G which satisfy the following condition

$$\sigma(T_i^m, T_j^n) > \varepsilon \; AND \; \big((n = m) \; OR \; (n = m + 1)\big)$$

Consequently, D contains only those nodes of G, which have at least one ε-related edge incident on it. In other words B contains all topic nodes that are ε-related to at least one more topic within the same year or across consecutive years.

6. For each edge in B, the weight $\sigma(T_i, T_j)$ is now recomputed as follows:

$$\sigma(T_i, T_j) = \sigma(T_i, T_j) / \nu \qquad (4)$$

where ν is the maximum of degrees of T_i and T_j. This reduces the weight of those edges that are connected to nodes which in turn are ε-related to many other nodes. Topics that represent generic and basic areas may overlap with many areas. Edges emanating from these topics get less priority. This step helps in suppressing noisy and obvious evolutions while giving priority to area-specific evolutions.

7. Arrange edges of B in decreasing order of associated weight $\sigma(T_i, T_j)$.
8. Remove the first element of B and initialize a cluster C with this element.
9. Repeat steps a to d until B is empty
   (a) Remove the top-most element ***e*** of B.
   (b) Add ***e*** to an existing cluster X of C if its addition maintains the κ connectivity in C. If ***e*** satisfies this relation with more than one cluster of C, add ***e*** to all such clusters.
   (c) Otherwise start a new cluster C'.
   (d) Update clusters $C = C \cup C'$
10. Output C.

The output of the above algorithm is a graph of connected components, where each component is a layered graph. A visualization of the graph is generated in which each layer is assigned a unique color. The layers are then presented in terms of increasing index of years from left to right.

Figure 3 illustrates two independent clusters from the topic evolution graph that was generated using all topics extracted from publications from 2007 to 2012. The bigger cluster shows the relationship of the areas Natural Language Processing (NLP), semantic web, gaming systems, online learning systems and social networks. This is obviously a correct and interesting evolution. It illustrates the continuing and important applications of natural language processing techniques to game-based learning and intelligent tutoring systems. The second cluster in figure 4, lower right corner shows continuing interest in support vector machines as a stand-alone topic.

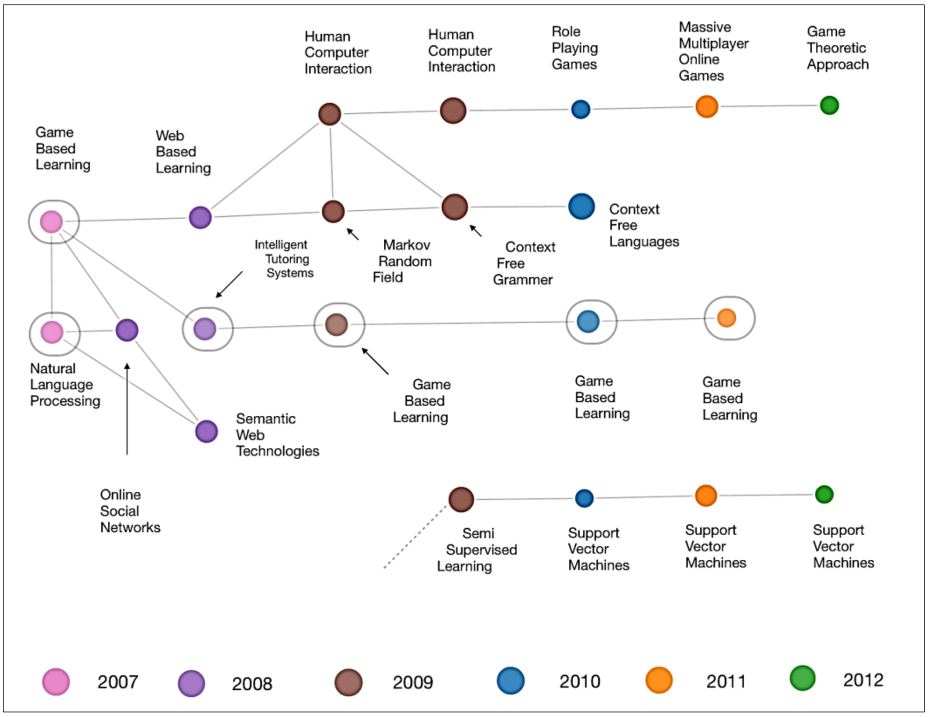


Fig. 3. Topic Evolution Graph (partial)

Table 1 summarizes the evolution history for a few popular topical phrases, using our method and c-ITM proposed in [17]. Most of these phrases were found by [17]. Columns 2 and 3 show the top phrases in predecessor and related topics of same or later years identified by the proposed method. Column 4 shows the topic names given for related topics as presented in [17]. Column 5 shows manual judgment about the relationship between these human-assigned topic names given by c-ITM and the topic phrases yielded by the proposed method.

Table 1: Topic Evolution - comparing proposed method with c-ITM

Top topic phrases and the year	**Topical phrases of Predecessor Topics (proposed method)**	**Topical Phrases of Related topics - Contemporary or Later (proposed method)**	**Predecessor Topic as per c-ITM [17]**	**Our observation**
ad hoc networks (2000)	Intersymbol interference isi(1995), Ahn collision detection(2000), Multicast rout-	Vehicular ad hoc(2006), Heuristics analytics system(2007)	Network communication since 1994	All phrases in columns 2 and 3 related to Network Communication

	ing protocols(2000)			
wireless sensor (2002)	Wavelength division multiplexing(2001) , Bluetooth 1.1 (2001), Sensor network systems(2002)	Underwater Sensor networks(2006) , Ubiquitous computing technologies(2007) , IP multimedia subsystem(2007)	sensor networks since 2003	Sensor Networks as a phrase was detected in 2002. Lot of related phrases were detected in 2001.
content based image retrieval (1995)	Markov random fields(1995) , Optical flow fields(1995)	Shear warp algorithm(2002) , Remote sensing image(2007), Context intelligent diagnosis(2009) , Medical image segmentation(2011)	hidden in information retrieval from 1993	Topical phrases in column 2 show evolution from Image Processing, Graphics and Hidden Information Retrieval
intrusion detection (2002)	Virtual private networks(2001), Denial service attacks(2001)	Access control policies(2003) ,	protocol security since 2000	Topical phrases are related to Protocol Security
support vector(2001)	Neural networks(2000) , Self organizing map(2000) , Principal component analysis(2000)	Hidden markov models(2002), Facial expression recognition(2006)	neural network since 2000	Topical phrases show evolution from Neural Networks
semantic web (2004)	Xml powered web(2003) , Web usage mining(2003)	Intelligent tutoring system(2005) , Service oriented computing(2006) , Web ontology language(2007), Formal concept analysis(2007) , Social text Streams(2007)	evolved from knowledge ontology since 2002	Topical phrases of predecessor and related topics depict significance of web ontology and xml based web architecture to semantic web
signature scheme (1995)	Public key infrastructure(1995) , Role based Access(1995) , Access control mechanisms(1995)	Buffer overflow attacks(2003) , Stolen verifier attack(2003), Key management system(2006)	protocol security since 2004	Significant methods/technologies related to protocol security emerge through Topical phrases of predecessor and related topics
fading channels(2000)	Code-division multiple access(1995), Bit error rate(1995),	Multiple access interference(2002)	channel coding since 2004	Topical phrases show evolution from channel coding

	Channel Impulse response(2000)			
xml data (2000)	Synchronized multimedia integration(2000)	Jsp xml web(2002) , Database management systems(2005) , Nearest neighbour queries(2005)	evolved from database since 2003	Role of database in the emergence of xml data formats is visible
energy consumption(2007)	Wireless sensor networks(2007), Sensor network applications(2007)	Dynamic voltage scaling(2007) , Pervasive computing environments(2008) , Energy harvesting systems(2010)	N/A	

4 Computing Commercialization Score of Topics

We now present a method to compute and present to the end-user a comprehensive view about the current state of commercialization of a research topic based on the patent volumes and patent trends applied in the area. Each topic is assigned an aggregate commercialization score based on its strength in an associated collection of patent applications. Presently, we have considered all patent applications that have been filed and/or granted with USPTO during the period of 2005 to 2013. However, the proposed method is generic and applicable for any collection.

Patents are also time-stamped documents. Each patent document is first subjected to phrase extraction. All 2-grams and 3-grams are extracted and used for indexing the patents. The Lucene indexer is used for the purpose of indexing and retrieving patent documents for a given topic.

Let T_i be a research topic belonging to the year y generated from publication analysis. Let S_i be the set of n topical phrases representing T_i. Let $\Psi(T_i)$ represent the commercialization score of T_i which is computed using an aggregated relevance score of the documents that are retrieved by Lucene for phrases in S_i as follows.

Let P_i denote all patent documents that contain at least one phrase from S_i. A document is said to contain a phrase if all the words of the phrase are found to lie within a window of w words in the document.

For each document retrieved by Lucene $d_i \in P_i$ relevance of d_i to topic T_i, $R(d_i, T_i)$ is computed as follows

$$R(d_i, T_i) = pFactor(T_i, d_i) * tNorm(T_i) * \sum_{p \in S_i}(f(p, d_i)^2 * I(p)^2 * rank(p))$$

where $f(p, d_i)$ term frequency of p in d_i and $I(p)$ is the inverse-document frequency,

$rank(p)$ is normalized significance of phrase p where the most significant phrase in S_i has maximum significance, pFactor(T_i, d_i) is the normalized score based on how many phrases of S_i are found in d_i, where the document that contains most topical phrases receives maximum weights,

tNorm(T_i) is a normalizing score computed as follows:

$$tNorm(T_i) = \frac{1}{\sqrt{sumOfSquaredWeights}}, where\ sumOfSquaredWeights = \sum_{p \in S_i} I(p) \quad (5)$$

A document may contain phrases belonging to more than one topic, though its relevance to each topic may differ. tNorm ensures that the document is considered more relevant to a topic T_x (say) than another topic T_y (say), if it contains highly significant phrases of T_x but less significant phrases of T_y.

Finally $\Psi(T_i)$ is computed as the logarithmic transform of the aggregated relevance scores of all documents containing topical phrases of T_i as

$$\Psi(T_i) = \log(\sum_{d_i \in P_i} R(d_i, T_i)). \quad (6)$$

The commercialization score is further discretized into a 5 point scale, using equal discretization over all non-zero scores, and are denoted by VERY HIGH, HIGH, MEDIUM, LOW and VERY LOW. Figure 4 presents a heat-map that illustrates the extent of commercialization for each of 100 research topics of 2012. The text pop-up shows that the topic of wireless-sensor networks has been heavily commercialized.

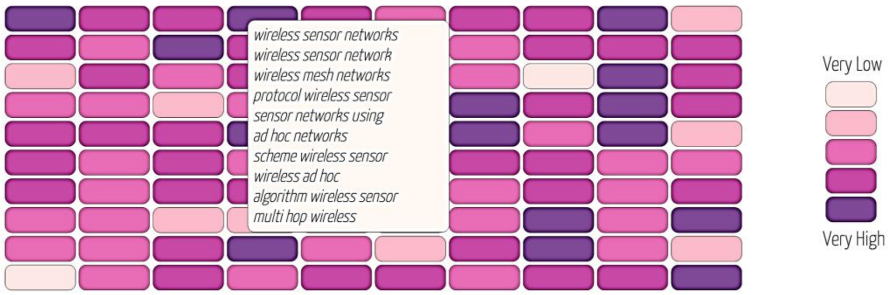


Fig. 4. Heat map showing aggregated commercialization of research topics of 2012. Wireless sensor Networks have been heavily commercialized

Figure 5 presents a graph, each of whose nodes are topics that represent the area of Wireless Sensor Networks, which is the same as the topic-evolution graph component for the area, with one difference. The size of a node in this graph is proportional to its commercialization score. The number of nodes in a particular year is indicative of the diversity of the topic as a whole. This graph also depicts that interest to file patents in this area had reached its peak in 2010.

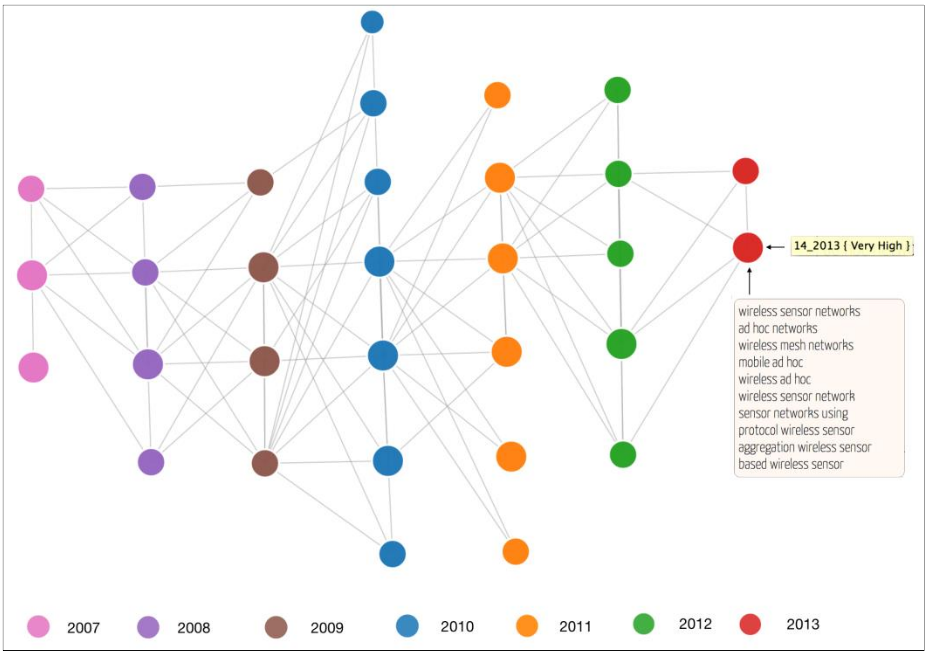


Fig. 5. Commercialization **trends of topics related to "Wireless Sensor Networks"**.

4.1 Analyzing Commercialization Trends

The history of commercialization of research topics can further lead to an understanding and categorization of commercialization of research areas into emerging, receding or yet-to-be-explored for potential commercialization. In order to detect trends, for a particular topic, say T_i^m, for a given year *m*, we first find all topics of past years that are maximally related to T_i^m using the topic evolution graph. Let L_i^m denote this list. Year-wise commercialization score for T_i^m is then computed using aggregate commercialization scores for all topics in L_i^m with the document collection restricted to those patent applications that have been filed in the year *m* only. Thus the yearly commercialization score for a topic T_i^m is given by

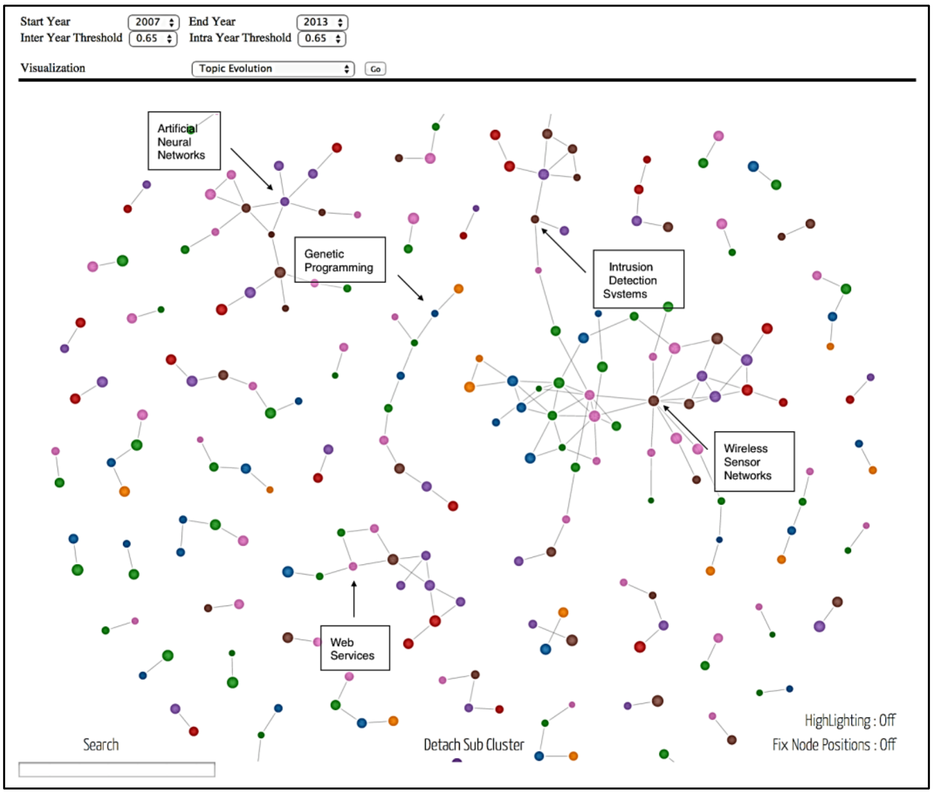
$$C(T_i^m) = \sum_{d_i^m \in P_i^m,\ t \in L_i^m} R(d_i^m, t) \tag{7}$$

where P_i^m denotes the collection of patent applications that have been filed in the year *m* and contains at least one topical phrase from the topics in L_i^m.

The total commercialization score along with trends of yearly commercialization scores are used for insight generation.

5 Experiments and Results

In this section we present some results from an implementation of the proposed methods to design a search system. The system has been implemented over a SOLR[1] based platform as a web-service. Research abstracts for the purpose were collected from sites dl.acm.org and csxstatic.ist.psu.edu/about/data, which have been made available by ACM and Citeseer respectively. The collection contains abstracts of Computer Science related publications along with title of paper, authors, venue and date of publication. After crawling, cleaning and indexing, the data has been stored locally on a server. All the proposed analytical methods run off-line to generate the similarity matrices and commercialization scores. Users can access the system as a web-application to search, drill-down and also see visualizations of topic evolution, commercialization etc. through appropriate inter-active visualizations.



Start Year 2007
End Year 2013
Inter Year Threshold 0.65
Intra Year Threshold 0.65
Visualization Topic Evolution

HighLighting : Off
Fix Node Positions : Off

Fig. 6. Topic Evolution Graph (Partial)

[1] Lucene.apache.org/solr

Figure 6 presents a few components of the topic evolution graph generated from all the topics for all years. Few nodes from two components have been highlighted. The first component on the left shows how areas of machine-learning and biological data mining have interacted over the areas. Similarly the component on the right shows that wireless technologies and privacy and security related research have influenced each other.

Table 2 presents the most commercialized research topics yielded by the system using the proposed commercialization scores, where research topics are extracted from the research publications. Table 3 presents some actionable insights generated from analysis of commercialization trends as stated at the end of section 4. On extreme left the column shows research areas that are hot, have commercial potential and not yet fully exploited. The second column indicates areas which are well-established and commercialization is on the rise. The third column shows areas that are very well-explored and saturated with patents and thus may be highly competitive to enter at this point. The fourth column shows areas which are theoretically well-explored and show declining trend of patenting.

Table 2. Top 10 most commercialized research topics (2005 - 2013)

Top 10 most Commercialized topics (2007 - 2013)
Using Mobile Devices
Cryptography
Wireless Sensor Networks
Real Time Systems
Image retrieval
Brain Computer Interface
Predictive Control for Autonomous Vehicle
Embedded Systems
Reduced Power Consumption
Intrusion Detection System

Table 3. Insights generated from Analysis of commercial Trends

New Research Areas - Very Few Patents - Rising Patent Trend	**Hot Research Areas - Many Patents - Rising Patent Trend**	**Popular research area – Large number of patents - Steady Patent Trend**	**Receding Research Area - Many patents - Patent Trend Decreasing**
Wheeled Mobile Robot	Multi-agent Systems	Wireless Sensor Networks	Collaborative Filtering
Human Robot Interaction	Support Vector Machines	Semantic Web	Service Oriented Architecture
	Social Network Analysis	Time Series Classification	
	Artificial Neural Network	Error Correcting Codes	

	Magnetic Resonance Imaging	Information Security	
	Cyber Physical Systems	Commercial Cloud Services	
	Electronic Health records		

Figure 7 presents the list of top 20 companies in USPTO database which have filed maximum patents in the areas listed in Table 2 between 2005 and 2013 along with the number of patents filed by them. Figure 8 (left) presents the most frequently occurring 3-gram phrases in patent applications for top 3 companies. On the right it presents phrases from patents by 3 companies which have filed a large number of patents in the areas listed in Table 1 only, though do not appear in the list of Figure 7. This shows an interesting aspect of commercialization. These are niche companies filing patents in specific trending areas of research. The established companies have a more diverse portfolio which includes many well-explored areas of research.

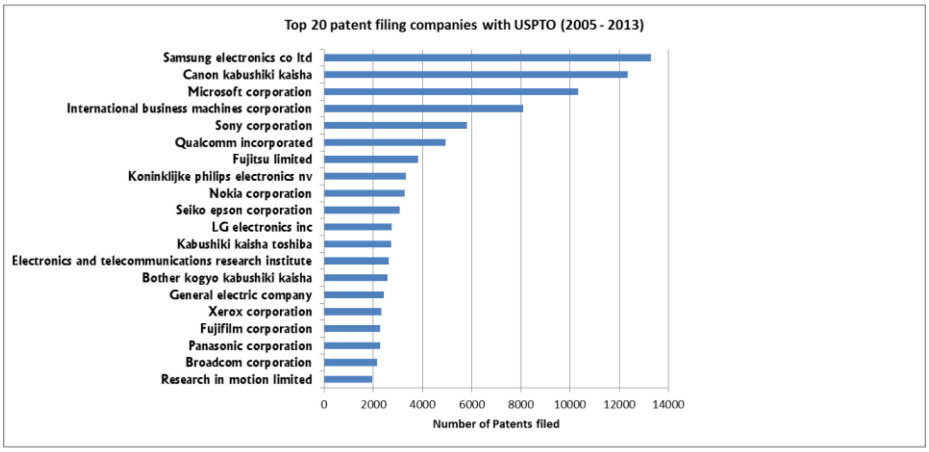


Fig. 7. Top 10 Companies filing patents in the above areas

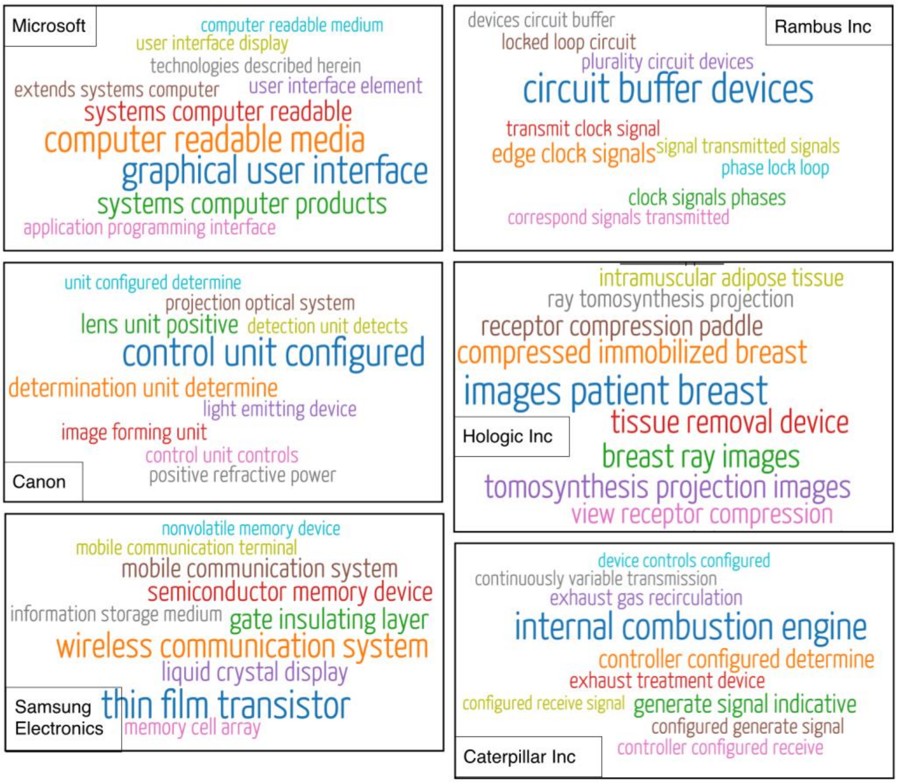

Microsoft
computer readable medium
user interface display
technologies described herein
extends systems computer
user interface element
systems computer readable
computer readable media
graphical user interface
systems computer products
application programming interface
Rambus Inc
devices circuit buffer
locked loop circuit
plurality circuit devices
circuit buffer devices
transmit clock signal
edge clock signals
signal transmitted signals
phase lock loop
clock signals phases
correspond signals transmitted
unit configured determine
projection optical system
lens unit positive
detection unit detects
control unit configured
determination unit determine
light emitting device
image forming unit
control unit controls
positive refractive power
Canon
intramuscular adipose tissue
ray tomosynthesis projection
receptor compression paddle
compressed immobilized breast
images patient breast
Hologic Inc
tissue removal device
breast ray images
tomosynthesis projection images
view receptor compression
nonvolatile memory device
mobile communication terminal
mobile communication system
semiconductor memory device
information storage medium
gate insulating layer
wireless communication system
liquid crystal display
Samsung Electronics
thin film transistor
memory cell array
device controls configured
continuously variable transmission
exhaust gas recirculation
internal combustion engine
controller configured determine
exhaust treatment device
configured receive signal
generate signal indicative
configured generate signal
Caterpillar Inc
controller configured receive

Fig. 8: Patent profile of companies through frequent phrases

6 Conclusions

In this paper, we have presented methodologies for analysing large volumes of research publications for information gathering and insight generation. We have presented results from an instance of implementation which currently analyses hundreds of thousands of research abstracts and patent applications jointly. The objective of the joint analysis is to come up with insights about current states of commercialization of research areas. Such a system helps in understanding current state of research as well as look for new ideas of commercialization. It also helps in understanding the existing competition.

Our future work lies in complete automation of the decision making process by aligning the content with external hierarchical indexing mechanisms like Wikipedia, journal content hierarchy etc. to explore inter-disciplinary topical relationships. This will help in better understanding of application of research areas and technologies to different areas for better decision making purposes.

References

1. Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper col-lections: Integrating statistics, text ana-lytics, and visualization. Journal of the American Society for Information Sci-ence and Technology, 63(12), 2351-2369.
2. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data min-ing (pp. 990-998).
3. Osborne, F., & Motta, E. (2012). Min-ing semantic relations between research areas. In The Semantic Web–ISWC 2012 (pp. 410-426). Springer Berlin Heidelberg.
4. Motta, E., & Osborne, F. (2012). Making Sense of Research with Rexplore. In 11th International Semantic Web Conference ISWC 2012 (p. 49).
5. Shahaf, D., Guestrin, C., & Horvitz, E. (2012, August). Metro maps of science. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data min-ing (pp. 1122-1130).
6. Yan, E., Ding, Y., Milojević, S., Sugimoto, C.R.: Topics in dynamic re-search communities: An exploratory study for the field of information retrieval. Journal of Informetrics, 6(1), 140-153. (2012)
7. Osborne, F., Scavo, G., & Motta, E. (2014). Identifying diachronic topic-based research communities by clustering shared research trajectories. In The Semantic Web: Trends and Challenges (pp. 114-129). Springer International Publishing.
8. Bozeman, B., Fay, D., & Slade, C. P. (2013). Research collaboration in universities and academic entrepreneurship: the-state-of-the-art. The Journal of Technology Transfer, 38(1), 1-67.
9. Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Pa-tent Information, 37, 3-13.
10. Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan and others (2012). Patentminer: topic-driven patent analysis and mining. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1366-1374.
11. Shih, M. J., Liu, D. R., & Hsu, M. L. (2010). Discovering competitive intelligence by mining changes in patent trends. Expert Systems with Applications, 37(4), 2882-2890.
12. Park, H., Kim, K., Choi, S., & Yoon, J. (2013). A patent intelligence system for strategic technology planning. Expert Systems with Applications, 40(7), 2373-2390.
13. Choi, S., Park, H., Kang, D., Lee, J. Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. Expert Systems with Applications, 39(13), 11443-11455.
14. David M. Blei, Andrew Y. Ng, and Mi-chael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Re-search, 3:993–1022, 2003.
15. T. L. Griffiths, M. Steyvers, D. Blei, and J. B. Tenenbaum. Integrating topics and syntax. Advances in Neural In-formation Processing Systems (2005).
16. Dey, L., Mahajan, D., & Gupta, H. (2014). Obtaining Technology Insights from Large and Heterogeneous Document Collections. In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on (Vol. 1, pp. 102-109).
17. Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? In CIKM, 2009.