RJ10450 (A0908-017) August 24, 2009
Other

IBM Research Report

SIMPLE: A Strategic Information Mining Platform for IP Excellence

Ying Chen, Scott Spangler, Jeffrey Kreulen, Stephen Boyer, Thomas D. Griffin, Alfredo Alba, Amit Behal, Bin He, Linda Kato, Ana Lelescu, Xian Wu*, Li Zhang*, Cheryl Kieliszewski
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
USA

*IBM Research Division
China Research Laboratory
Building 19, Zhouguancun Software Park
8 Dongbeiwang West Road, Haidian District
Beijing, 100193
P.R.China

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

SIMPLE: A Strategic Information Mining PLatform for IP Excellence

Ying Chen, Scott Spangler, Jeffrey Kreulen, Stephen Boyer, Thomas D Griffin, Alfredo Alba, Amit Behal, Bin He, Linda Kato, Ana Lelescu, Xian Wu[1], Li Zhang[1], Cheryl Kieliszewski
{yingchen, sboyer, tdg, aalba, abehal, binhe, kato, alelescu, cher}@us.ibm.com,
{spangles, kreulen}@almaden.ibm.com
{wuxian, lizhang}@cn.ibm.com
IBM Almaden Research Center, San Jose, CA 95120
[1]IBM China Research Center, Beijing, China

Abstract

Intellectual Properties (IP), such as patents and trademarks, are one of the most critical assets in today's enterprises and research organizations. They represent the core innovation and differentiators of an organization. When leveraged effectively, they not only protect a business from its competition, but also generate significant opportunities in licensing, execution, long term research and innovation. In certain industries, such as Pharmaceutical industry, patents lead to multi-billion dollar revenue each year. In this paper, we present a holistic information mining solution, called SIMPLE, which mines large corpus of patents and scientific literature for insights. Unlike much prior work that deals with specific aspects of analytics, SIMPLE is an integrated and end-to-end IP analytics solution which addresses a wide range of challenges in patent analytics such as the data complexity, scale, and nomenclature issues. It encompasses techniques and tools for patent data processing and modeling, analytics algorithms, web interface and web services for analytics service delivery and end-user interaction. We use real-world case studies to demonstrate the effectiveness of SIMPLE.

1. Introduction

With the explosion of diverse types of information in organizations and in public, text and data mining solutions are now receiving unprecedented attention. Text and data mining solutions analyze large volume of unstructured and structured data respectively to bring insights to users. In particular, Intellectual Properties (IP) represents one of the most valuable information assets to corporations. Appropriate management and leverage of IP information can create significant competitive advantages, generate high-value and low-cost returns through licensing and divesting opportunities, and enable major science and technology breakthroughs. Today, some industries rely primarily on the IP related activities and business to survive and thrive, such as healthcare and life sciences and pharmaceutical industries.

IP activities may range from prior art search, portfolio analysis and management, licensing target identification, divestiture analysis, to patent valuation. In specific industry, semantic entity extraction from patents is important as well. For instance, extracting chemical names and biological entities from patents is crucial for drug research. To date, many such IP activities rely on tedious, expensive and error-prone manual processing. Machine-aided analytics are becoming increasingly essential. Yet the large variation in the quantity, quality and unique characteristics of IP information makes it especially challenging for adopting many exiting text and data mining solutions as is. For instance, a typical IP data corpus for Pharmaceutical research may comprise collections of granted patents and applications from US, European, and World-wide patent offices, scientific literature such as PubMed Medline scientific articles [1] as well as other raw data produced by high throughput screening. This patent corpus alone may contain over 10 million documents and consumes hundreds of Gigabytes of storage and it is continuously updated with new data. An end-to-end IP analytics solution must take into consideration issues related to data volume, diversity, and speed of change and it must include a wide range of data processing and analytics tools to derive insights.

Mining patents requires addressing three major technical challenges. First, a solution requires management of the information itself. Processing, cleansing, normalizing, validating and storing the large volume of information in a manner that it is ready and accessible for downstream analysis. Given the unique characteristics and legal significance of patent data, this step is especially critical. Second, we need to apply interactive and batch analytical techniques to the structured and unstructured information to derive additional value-added attributes and relationships. These techniques consist of technologies such as machine learning [2], clustering and classification [3, 4, 5], and entity extraction (also called annotation) algorithms using Natural Language Processing or otherwise [6, 7, 8]. Third, it requires transformation of the information into a human interactive interface and consumable form such as reports and visualizations.

In this paper, we present such a holistic IP mining solution called SIMPLE. SIMPLE consists of a suite of tools and processes for processing IP data and data warehousing, a set of analytics technologies and tools for patent analysis, a web-service enablement of the analytical services in a service-oriented architecture (SOA), and a web based user interface and visualizations for end user consumption of analytical results. SIMPLE has been successfully used in many real-world scenarios. In the rest of the paper, we present the key challenges in mining patent data in Section 2. We then present the overall SIMPLE system architecture and its key components in Section 3. Several real-world case studies are illustrated in Section 4. Finally, Section 5 concludes and outlines the future work.

2. Patent mining challenges

Analyzing patents is particularly challenging. First, raw patent data provided by different authorities is widely available in different formats, e.g., XML [9] and images. However, such raw data is complex. Patents contain a large set of highly valuable structured fields such as inventors, assignees, dates, and class codes. Yet they are often not normalized or standardized. For example, an assignee may have many different assignee name variations. Without appropriate normalization, searching for a specific assignee name can only result in a subset of patents rather than the entire patent portfolio from that assignee. As for unstructured text fields such as title, abstract, claims, and text body of the patents, they often contain various encodings that present roadblocks for parsing, search, and text analysis.

The challenge is exacerbated when dealing with pharmaceutical specific analysis, such as searching chemical names and biological entities from patents. This is because the nomenclature associated with chemical substances is difficult to understand. Inconsistencies among chemical terms are widespread, despite the standard efforts. For example, “Valium” has at least 149 other names. Many can be found in patents. Searching for certain pharmaceuticals in the patent literature using commonly accepted phrases is extremely difficult. Another form of nomenclature challenge has to do with the patent language, especially claims. Patent claims are the heart of the patents. The ability to analyze claims language to derive insight is essential to a successful IP analytics solution. In our experience, we found that new algorithms and analytical techniques are often required. We will highlight a few such algorithms in this paper.

Finally, the size of the overall patent corpus poses a different dimension of challenge for research and development activities. Today, running a chemical name annotator against the entire patent corpus would take weeks on a sizable server. When parallelized on clusters or the Blue Gene supercomputer, it may take only minutes. Clearly, scalable algorithms and technologies are needed when handling such large data corpus.

3. SIMPLE system architecture

To address the challenges above, SIMPLE embeds four major components that are integrated in an end to end fashion. The overall system architecture is shown in Figure 1. The four key components are the follows:

1. A **General Extract-Transform-Load (GETL) engine** which processes the raw patent data into a clean data warehouse containing both structured and unstructured text information. Such a tool can also be used to process content other than patents, e.g., Medline scientific articles and web pages. It also generates appropriate indices for general keyword based text and structured data search using existing search engines such as Lucene [10].
2. **Annotators, such as chemical and biological entity annotators,** that extract semantic entities, e.g., chemical names drugs, diseases, and Genes from unstructured text. The extracted entities are stored back into the data warehouse as additional structured data to enable subsequent search and analysis.
3. A run-time **analytics engine** that performs different types of runtime analytics. For example, Nearest Neighbor search (NN) for searching for prior art, patent Claims Originality analysis (CO) for ranking patents, and patent clustering (PC) for portfolio analysis, and relationship analysis for finding relationships of multiple dimensions of information. Such analytics are provided via web services. This enables other applications to integrate with SIMPLE analytics services easily.
4. **A web based user interface and a set of visualization components** for user consumption of the analytical results.

Overall SIMPLE’s GETL engine ingests the patent feeds on an ongoing basis, extracts desired information from them, transforms and cleanses them into standard formats, and loads them into the data warehouse modeled by a standard data warehouse model, e.g., star or snowflake schema [11, 12]. The data warehouse enables

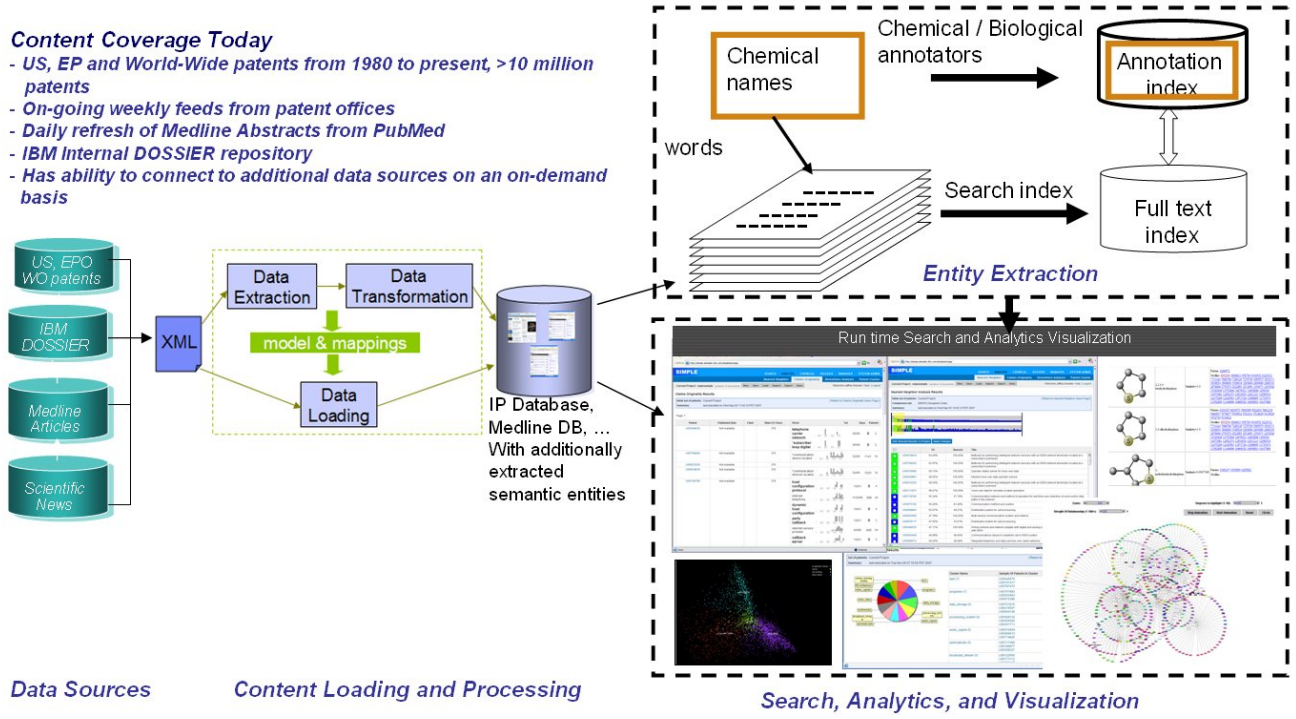


Figure 1. Overall SIMPLE system architecture.

efficient Business Intelligence-type aggregations and reporting. The cleansed text data in the data warehouse can then be fed into a suite of annotators to derive additional semantic entities from the text. The extracted entities are inserted back into the data warehouse as additional structured dimensions for subsequent online analysis. Once the data warehouse is fully populated, runtime search and analytics are applied to derive insights. The analytical results are shown via a web interface with visualizations. Below we briefly describe the key functions of each of the four major system components.

3.1 GETL for Patent Processing

To construct and maintain the ongoing patent data feed into a cleansed data warehouse, we developed the GETL engine (see Figure 2 for its system architecture). It includes four key sets of technologies as discussed below (see [13] for the details of the GETL approach).

1. A general data model mapping framework that maps arbitrary XML data sources to standard data warehouse data models such as star and snowflake schema.
2. A general data Extraction, Transformation, and Loading framework that allows flexible extraction of various fields in the source data, diverse transformation of source fields into user- or system-defined formats in the target data warehouse, and efficient loading of large amount of data into database tables.
3. A set of cleansing capabilities such as duplicate detection and elimination and filtering of data by specific fields of user-defined functions during data loading.
4. A set of recovery capabilities that allow undo, redo, abort and recovery operations in the face of data loading errors and failures.

3.2 Chemical name and Biological entities annotators

Annotation technologies entail processing of text information and extracting the desired entities out of it through various forms of analytical processing. SIMPLE currently includes a set of chemical and biological name annotators. Chemical annotation is a multi-step process. First we applied a set of analytics technologies to extract candidate chemical names from text. All the extracted potential chemicals are then fed into a name-to-structure converter program that converts chemical names into structures. Such a program makes no value judgments, focusing only on providing a structure that the name accurately describes [14]. Remarkably, this process also serves as a 'normalization' process that maps chemicals that have different names to the same structure,

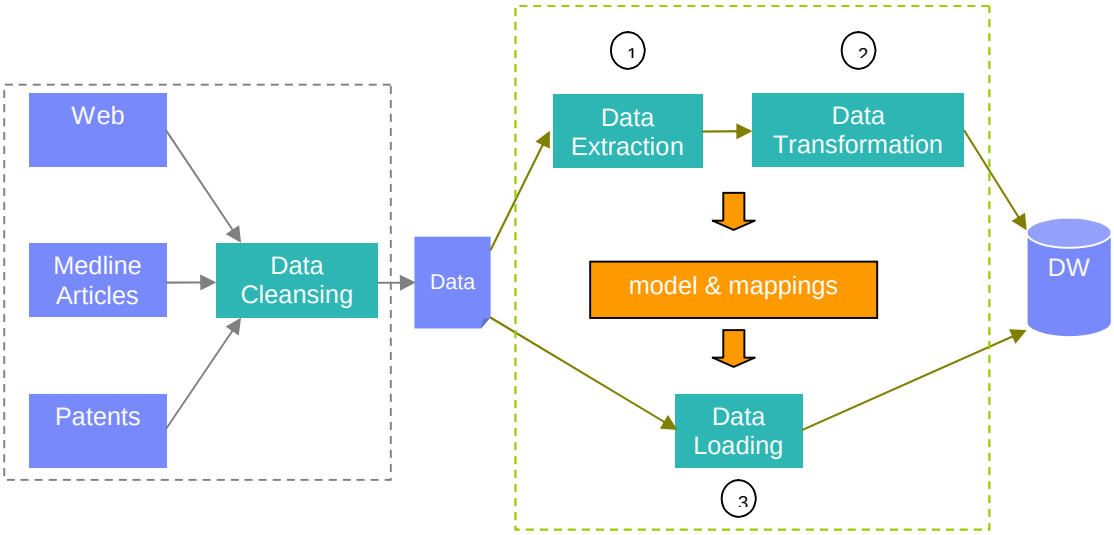


Figure 2. GETL engine architecture.

represented by a common and standard SMILES string [15].

While developing the chemical name annotator, we also experimented with multiple annotation technologies, ranging from hidden Markov (HMM) models [16], conditional random field theory (CRF) [17], to dictionary and rule-based technologies. For chemical names, we found that a combination of rule-based and negative dictionary based approaches work particularly well. This is because chemical names are typically 'unlike' other words commonly used in the English language. By appropriately applying dictionary technologies and domain-knowledge-driven rules, we can devise high quality annotators (see [18] for details of the SIMPLE chemical annotator technology). Using such techniques, we were able to identify close to 140 million chemical names from over 30 years of US Patent corpus and close to 12 million from Medline scientific article abstracts.

Annotation runs on patents often take several weeks to complete. To speed up this process, we devised a set of analytical approaches to allow annotation processes to run on massively parallel supercomputers, such as IBM's Bluegene [19]. Through such parallelization, we were able to reduce weeks of annotation time down to minutes. Such rapid annotation techniques open up possibilities for corporations to provide supercomputer-powered annotation services. For example, individual pharmaceutical companies may submit certain customizations to the annotator such as using their own proprietary dictionaries and request an annotator run for the entire patent and Medline data corpus and receive results within minutes. This will significantly speed up the entire drug discovery research and development activities.

3.3 Runtime Analysis and Visualization

With the cleansed data warehouse and the extracted new annotations, additional runtime analytics can be applied to derive insights. SIMPLE analytics are delivered via both a web service API and a web interface. The web service interface allows other applications to integrate with SIMPLE's analytics services. The web interface with visualizations allows the end user to consume the analytical results easily. We describe several such runtime analytics capabilities in SIMPLE below.

3.3.1 Nearest Neighbor Search (NN)

One of the most critical IP activities is collecting together a group of patents on a similar subject for potential licensing as a package. Such search intends to find a set of patents within a portfolio that are similar to a given set of known patents. To increase the value of such a licensing package, it is desirable to find additional similar patents that can be included in the package as well.

Given the input patents, the search result is sorted by the "closeness" of the result set to the given set such that the most similar ones are shown at the top of the list. In addition to using patents as inputs, NN may also take a short paragraph of text as input. In such cases, NN will try to find patents that have similar text. In addition to its value in finding patents for licensing, by using this text input feature NN is often used by patent attorneys to identify prior art for a proposed invention disclosure.

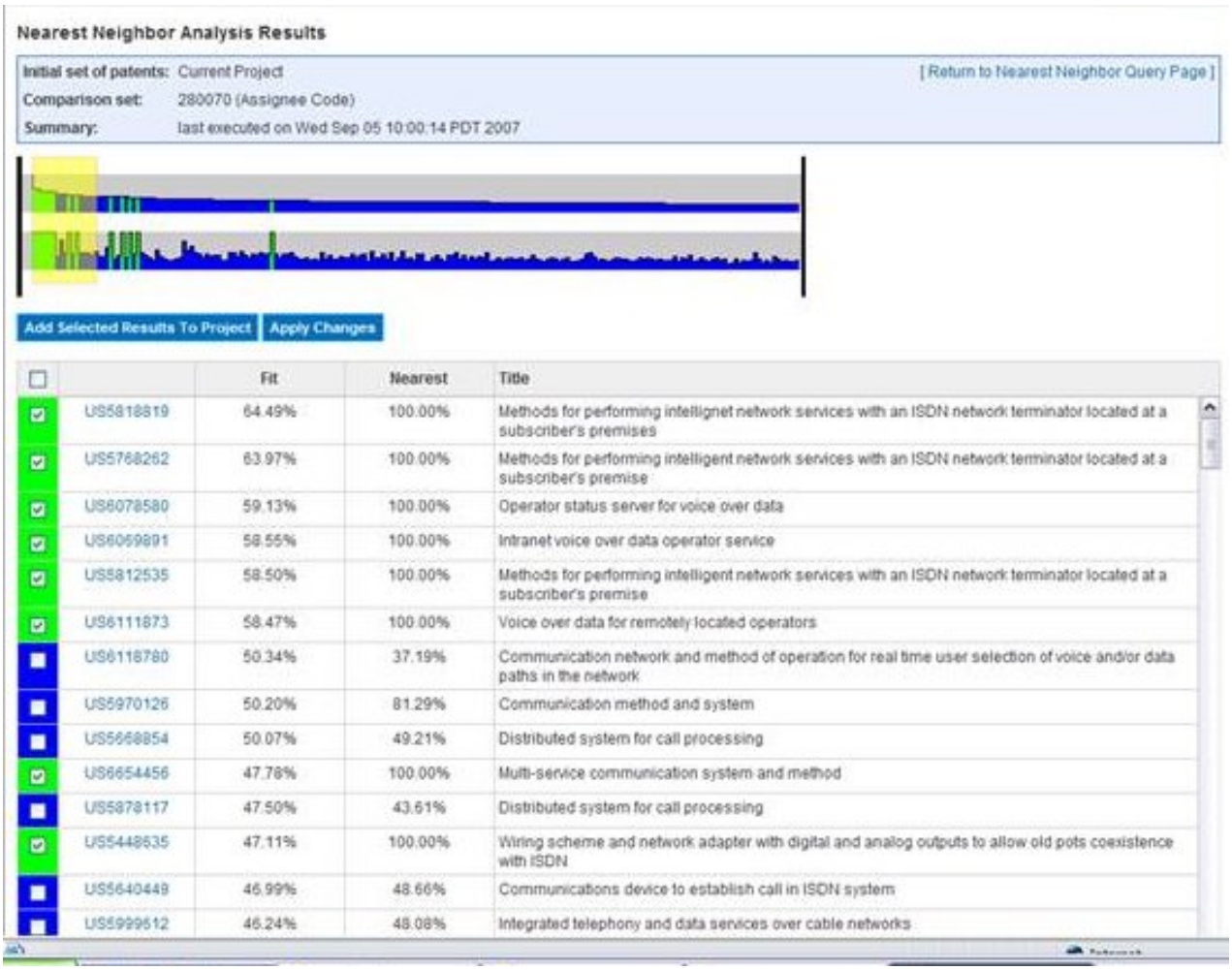


Figure 3. Nearest Neighbor Search Result Visualization.

SIMPLE's NN creates a text cluster based on the given set of patents. In such text clusters, each document is represented by a vector of words and phrases (also called a Vector Space Model) extracted from the patent text body. The overall text cluster is represented by a centriod model, modeled and computed by the vectors that represented the input documents. [20] provides a detailed description on text clustering, centriod models and vector space models. With such document modeling techniques, one can compute the distance between any two patents represented by the two vectors of features.

To compute the distance of the other patents to the input set, we measure the distance between the other candidate patents and the centroid of the cluster. The candidate pool of patents is drawn from all those in the portfolio that share membership in the same IPC (international patent code) classes as the original input patents. IPC classes are high level, standard, and broad patent classifications established by patent authorities.

Using IPC classes to form the candidate pool allows us to limit the universe that the NN has to search against without significantly compromising on the quality of the results. Meanwhile, limiting the search scope can improve the overall performance.

In SIMPLE, NN returns two sorted list of patents, one sorted by the distance between a result patent and the centriod of the input cluster (called "fitness" values) and the other sorted by the distance between the result patent and any individual patent in the input set (called "Nearest" values). The first list treats the input patents as a single cluster with a centriod. The second list can be useful if the input patents are not alike themselves. One may be interested in finding patents that are close to any of the patents in the given set instead of the overall set. Figure 3 shows an example of NN visualization output. The light-shaded patents are the inputs. The dark bars represent the resulting patents, sorted by the "fitness" value in the view.

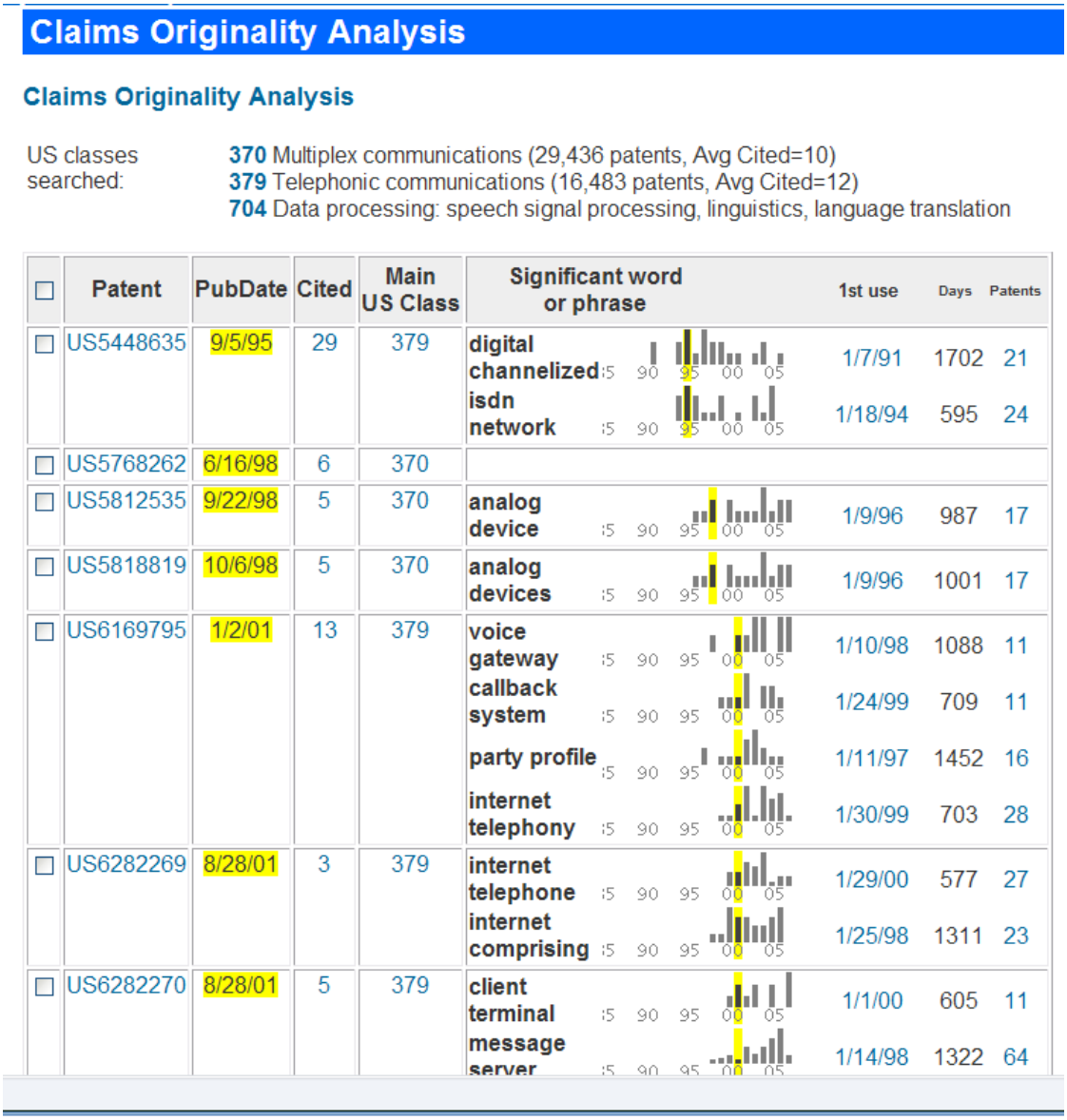


Figure 4. Claims Originality Analysis Visualization.

3.3.2 Claims Originality Analysis (CO)

The ability to rank or score the quality of patents is critical in many IP tasks. Claims Originality analysis intends to provide a form of ranking based on analysis of patent claims language. Specifically, it evaluates each patent's technical contribution to the field and brings to the notice of the analyst, those patents that are most valuable, and why they are deemed to be so. This approach helps highlight specific words and phrases in the patent claim section that help to demark the technical contribution of the patent and by aggregating and counting both the earliness and subsequent usage of these phrases helps to critically measure the potential licensing value of the patent in question (see [21] for the CO details). Figure 4 shows an example of the CO analysis result for a given set of patents.

Visually, the results are presented in a table format with the following columns: patent number, publish date, class-code, citation count, key phrases, and the rating value. For each of the identified key phrases, we also show the first use date, the day difference (inverse of recency) and the support value for that phrase, i.e., the unique number of

Class Name	Size	Trend /
client server architectures	962 (3.72%)	11/19/00
operating system	1248 (4.82%)	9/27/00
semiconductors	909 (3.51%)	9/12/00
database	598 (2.31%)	8/12/00
file systems	506 (1.95%)	8/8/00
text	216 (0.83%)	7/3/00
fabrication of chips	1761 (6.80%)	4/25/00
photoresist	624 (2.41%)	2/27/00
software architecture	915 (3.53%)	2/3/00
circuit	1877 (7.25%)	1/8/00
message handling	401 (1.55%)	1/2/00
films and surfaces	1814 (7.01%)	12/31/99
materials	1104 (4.26%)	12/24/99
Miscellaneous	2356 (9.10%)	12/4/99
networks	1171 (4.52%)	11/26/99
displays	852 (3.29%)	11/5/99
memory	1873 (7.23%)	10/27/99
processor instructions	668 (2.58%)	10/24/99
storage	913 (3.53%)	10/2/99
signal processing	1472 (5.68%)	9/20/99
image processing	667 (2.58%)	9/12/99
disk drives	1190 (4.60%)	8/15/99
chemical compounds	248 (0.96%)	8/3/99
computer device communication	513 (1.98%)	7/18/99
polymer	206 (0.80%)	5/18/99
video	830 (3.21%)	5/13/99
TOTAL / AVERAGE	25894	

Figure 5. IBM Patent Portfolio Summary (1994-2002) based on SIMPLE's Patent Clustering Analysis.

patents that subsequently used the same phrases. For instance, from the table, we notice that, while ranking patent 5448635, one of the important phrases is "isdn network", it was first used in a patent published in 1/18/1994, which is 595 days before this patent. The support of the term is 24 patents, i.e., after the first use, the term has been subsequently used in 24 distinct patents. The "first use" date of any phrase is hyper-linked to the text of the patent that used that phrase for the first time. Supports are also linked to all of the supporting patents.

3.3.3 Patent Cluster

Patent clustering generates patent clusters for a given set of patents based on the patent text fields, such as abstracts, claims, and titles. The generated patent clusters can form a taxonomy that categorizes the given patent set. Although patents typically come with different categorizations already based on the structured fields, such as IPC class code, such standard classification is often too broad or too high level. The text-based classification using patent clustering is more truthful and reflective, since it is based on patents' text content.

Patent clustering is a powerful tool for portfolio analysis. For example, when it is applied to the patents for a given corporation, the generated patent clusters naturally represent different categories of patents that the corporation might have. Our patent cluster also embeds additional information such as trends and statistical results of the clusters, e.g., cohension and distinctness values, to allow users to further understand the quality of the clustering results and cluster trends.

Figure 5 shows an example of patent clustering result based on IBM's patents between 1994 and 2002. The small trend lines indicate which categories of patents represent recent trends. For example, client-server architecture category had an upward trending, indicating that it might be an area that IBM is paying more attention to during late 1990's and early 2000's. On the contrary, video and polymer patents are trending downward, indicating that there may be less work going into those areas in that time period.

3.3.4 Relationship Analysis

Patent clustering-generated taxonomy, annotation results such as chemical and biological names and structured fields represent different dimensions of information. They allow users to analyze patents from different angles. In addition, SIMPLE's relationship analysis allows multiple dimensions of information to be analyzed at the same time, to identify interesting

Figure 6. The Relationship Analysis between chemical names and biological entities.

correlations. For example, when we perform relationship analysis on chemical names and medical conditions, we can find which medical conditions are highly related to which chemicals by analyzing patents whose claims mention the molecule as well as medical conditions.

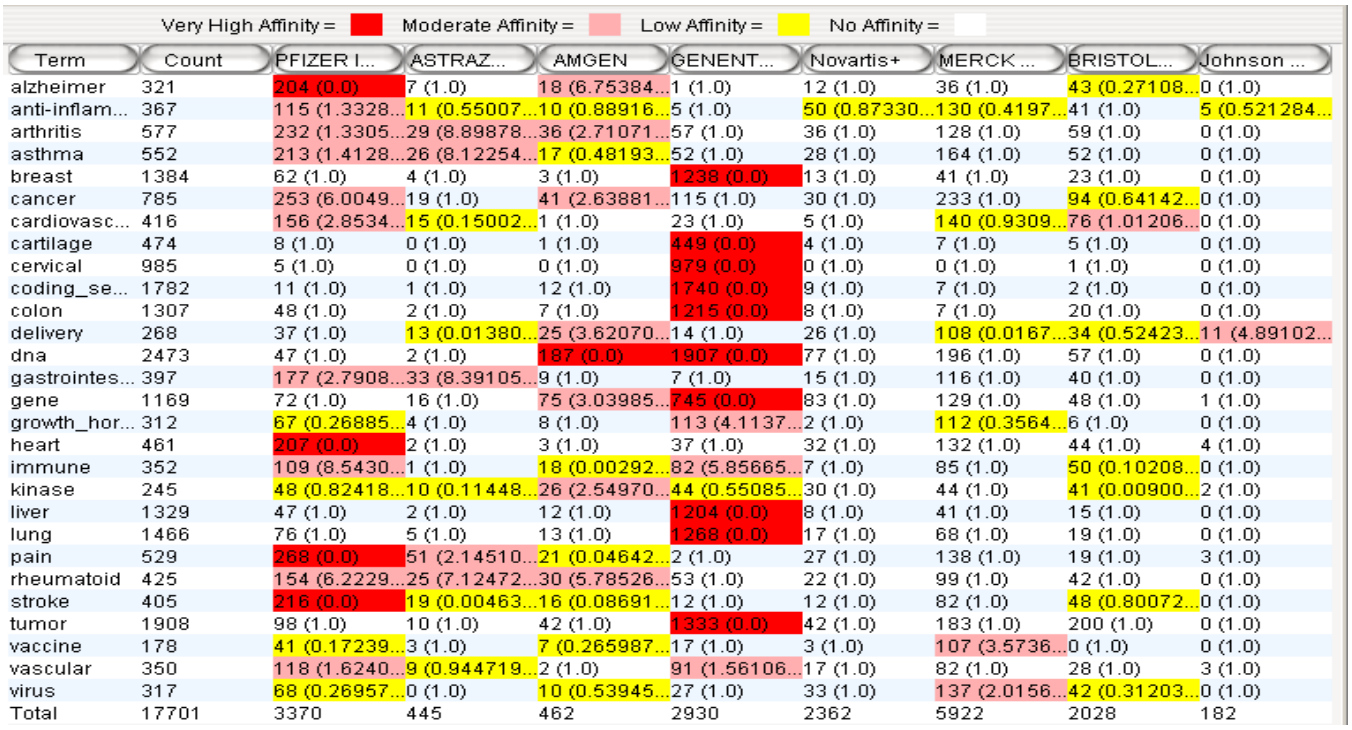
Logically, the molecule and a medical condition are considered to have high correlation/affinity if a large number of patents contain both entities in the claims. Specifically, SIMPLE uses a Chi-squared test is used to compute the affinity [22]. Figure 6 shows an example. For caffeine, migraine and headache have a high affinity, nausea and anxiety a moderate one, and burns and cough a low affinity.

Overall, such runtime analytics are embedded in an analysis workflow. The workflow may start with SIMPLE's patent search capability to identify a set of patents that are of interest. Then the set is fed into the analytics services. For example, NN would find other relevant patents for the given set. CO would be able to rank these patents in terms of the originality of the claims.

SIMPLE also contains other forms of runtime analysis such as patent divestiture analysis, emerging terms identification, and trends. Divestiture analysis can analyze the impact of divesting a set of given patents by reporting simulated changes after divestiture. Emerging terms analysis can identify words and phrases that are becoming more and more prominent for a given class of patents. We do not illustrate details of these analytics due to space limitation. To further assist users, users can save the search and analysis results into "projects". Projects can then be shared among groups, and exported into different formats.

4. Case studies

4.1 IBM Patent Licensing



Very High Affinity = Moderate Affinity = Low Affinity = No Affinity =

Term	Count	PFIZER I...	ASTRAZ...	AMGEN	GENENT...	Novartis+	MERCK ...	BRISTOL...	Johnson ...
alzheimer	321	204 (0.0)	7 (1.0)	18 (6.75384...	1 (1.0)	12 (1.0)	36 (1.0)	43 (0.27108...	0 (1.0)
anti-inflam...	367	115 (1.3328...	11 (0.55007...	10 (0.88916...	5 (1.0)	50 (0.87330...	130 (0.4197...	41 (1.0)	5 (0.521284...
arthritis	577	232 (1.3305...	29 (8.89878...	36 (2.71071...	57 (1.0)	36 (1.0)	128 (1.0)	59 (1.0)	0 (1.0)
asthma	552	213 (1.4128...	26 (8.12254...	17 (0.48193...	52 (1.0)	28 (1.0)	164 (1.0)	52 (1.0)	0 (1.0)
breast	1384	62 (1.0)	4 (1.0)	3 (1.0)	1238 (0.0)	13 (1.0)	41 (1.0)	23 (1.0)	0 (1.0)
cancer	785	253 (6.0049...	19 (1.0)	41 (2.63881...	115 (1.0)	30 (1.0)	233 (1.0)	94 (0.64142...	0 (1.0)
cardiovasc...	416	156 (2.8534...	15 (0.15002...	1 (1.0)	23 (1.0)	5 (1.0)	140 (0.9309...	76 (1.01206...	0 (1.0)
cartilage	474	8 (1.0)	0 (1.0)	1 (1.0)	449 (0.0)	4 (1.0)	7 (1.0)	5 (1.0)	0 (1.0)
cervical	985	5 (1.0)	0 (1.0)	0 (1.0)	979 (0.0)	0 (1.0)	0 (1.0)	1 (1.0)	0 (1.0)
coding_se...	1782	11 (1.0)	1 (1.0)	12 (1.0)	1740 (0.0)	9 (1.0)	7 (1.0)	2 (1.0)	0 (1.0)
colon	1307	48 (1.0)	2 (1.0)	7 (1.0)	1215 (0.0)	8 (1.0)	7 (1.0)	20 (1.0)	0 (1.0)
delivery	268	37 (1.0)	13 (0.01380...	25 (3.62070...	14 (1.0)	26 (1.0)	108 (0.0167...	34 (0.52423...	11 (4.89102...
dna	2473	47 (1.0)	2 (1.0)	187 (0.0)	1907 (0.0)	77 (1.0)	196 (1.0)	57 (1.0)	0 (1.0)
gastrointes...	397	177 (2.7908...	33 (8.39105...	9 (1.0)	7 (1.0)	15 (1.0)	116 (1.0)	40 (1.0)	0 (1.0)
gene	1169	72 (1.0)	16 (1.0)	75 (3.03985...	745 (0.0)	83 (1.0)	129 (1.0)	48 (1.0)	1 (1.0)
growth_hor...	312	67 (0.26885...	4 (1.0)	8 (1.0)	113 (4.1137...	2 (1.0)	112 (0.3564...	6 (1.0)	0 (1.0)
heart	461	207 (0.0)	2 (1.0)	3 (1.0)	37 (1.0)	32 (1.0)	132 (1.0)	44 (1.0)	4 (1.0)
immune	352	109 (8.5430...	1 (1.0)	18 (0.00292...	82 (5.85665...	7 (1.0)	85 (1.0)	50 (0.10208...	0 (1.0)
kinase	245	48 (0.82418...	10 (0.11448...	26 (2.54970...	44 (0.55085...	30 (1.0)	44 (1.0)	41 (0.00900...	2 (1.0)
liver	1329	47 (1.0)	2 (1.0)	12 (1.0)	1204 (0.0)	8 (1.0)	41 (1.0)	15 (1.0)	0 (1.0)
lung	1466	76 (1.0)	5 (1.0)	13 (1.0)	1268 (0.0)	17 (1.0)	68 (1.0)	19 (1.0)	0 (1.0)
pain	529	268 (0.0)	51 (2.14510...	21 (0.04642...	2 (1.0)	27 (1.0)	138 (1.0)	19 (1.0)	3 (1.0)
rheumatoid	425	154 (6.2229...	25 (7.12472...	30 (5.78526...	53 (1.0)	22 (1.0)	99 (1.0)	42 (1.0)	0 (1.0)
stroke	405	216 (0.0)	19 (0.00463...	16 (0.08691...	12 (1.0)	12 (1.0)	82 (1.0)	48 (0.80072...	0 (1.0)
tumor	1908	98 (1.0)	10 (1.0)	42 (1.0)	1333 (0.0)	42 (1.0)	183 (1.0)	200 (1.0)	0 (1.0)
vaccine	178	41 (0.17239...	3 (1.0)	7 (0.265987...	17 (1.0)	3 (1.0)	107 (3.5736...	0 (1.0)	0 (1.0)
vascular	350	118 (1.6240...	9 (0.944719...	2 (1.0)	91 (1.56106...	17 (1.0)	82 (1.0)	28 (1.0)	3 (1.0)
virus	317	68 (0.26957...	0 (1.0)	10 (0.53945...	27 (1.0)	33 (1.0)	137 (2.0156...	42 (0.31203...	0 (1.0)
Total	17701	3370	445	462	2930	2362	5922	2028	182

Figure 7. The Relationship Analysis between IP categories for the Pharmaceutical industry and companies.

One of the first utilizations of the SIMPLE application was in the IBM Trademark and Intellectual Property department. IBM owns more than 40,000 patents [23], in more than twenty different patent classes. One of the functions of this group is to find patents that IBM owns that are considered to be outside of our core business and sell or license them to other organizations. In one case, they gave us input to SIMPLE as set of 13 patents in the Voice over Internet Protocol space, looking to find additional patents that might be of value in this space. We ran SIMPLE's NN analysis on these input patents and found many more IBM patents that are relevant.

Once the nearest neighbor patents are identified, CO analysis was performed to evaluate the relative value of each of the top 10 identified patents. In the end, the IBM IP staff expanded the original patent package to 20 patents with an appropriate understanding of the relative value of the patents. This led to much more significant licensing revenue with a shorter processing time. There are several other similar cases. Due to space limitation, we do not present in this paper.

4.2 Industry-level patent landscape analysis

SIMPLE's patent clustering and relationship analysis have been used by IBM's Global Business Services in understanding IP landscape and trends for a given industry. For example, we have done an IP landscape analysis for Pharmaceutical industry by looking at US patent trends from 1980's to 2004. Our analysis showed who have increasing IP activities and who have decreasing ones.

In addition, we performed patent clustering on the patents from the entire industry and found emerging patterns as indicated by the patent clusters, such as Alzheimer, arthritis, etc. Once the patent clusters are generated, we performed relationship analysis between the generated categories and the assignee names, i.e., companies. Such an analysis allows us to correlate companies with their focus domains and identify whitespaces in the industry. Figure 7 shows the result of such analysis. The column headings indicate the companies. Each row corresponds to one cluster/category of patents. The shading indicates the level of affinity between the company and the corresponding category. The dark/red cell shows the highest affinity between the company and the corresponding category. As shown in Figure 7, Genentic is highly associated with the category of "tumor", "liver", "lung", etc. More interestingly, Genentic has the most unique patent portfolio. The areas that it focuses on are white spaces for other companies by and large (the rows that Genetic has dark cells on have no other dark cells in other columns) .

5. Conclusions and future work

In this paper, we presented a holistic analytics solution for mining patents and scientific literature. Our solution integrates ETL data processing, data modeling, offline analytics, e.g., chemical name annotation, runtime analytics, e.g., NN and CO, and end user visualization. The overall system is further scaled up by utilizing supercomputer resources, e.g., blue gene. In the future, we plan to experiment SIMPLE on the cloud infrastructure [24], as alternative low-cost and high performing platforms. We also plan to improve real-time performance of SIMPLE's runtime analytics, such as NN, CO, and patent clustering. Currently, such analysis takes minutes to run for a relatively small input patent set (e.g., tens of patents). New algorithms and techniques are needed to scale it up to larger inputs.

6. References

[1] PubMed Medline, http://www.ncbi.nlm.nih.gov/pubmed/.

[2] Alpaydın, E, Introduction to Machine Learning (Adaptive Computation and Machine Learning), *MIT Press*, ISBN 0262012111, 2004

[3] Spangler, W. S and Kreulen, J,, Mining the Talk: Unlocking the Business Value of Unstructured Information, *IBM Press*, 2007

[4] Modha, D., and Spangler, S.: Feature weighting in K-Means clustering. *Machine learning*. 52:3:217-237. 2003.

[5] Spangler, S., Kreulen, J., and Lesser, J.: Generating and browsing multiple taxonomies over a document collection. *J. of Management Information Systems*. Vol. 19. No. 4, pp 191-212. 2003.

[6] Gotz, T. and Suhre, O., Design and implementation of the UIMA common analysis system, *IBM System Journal,* Vol. 43, No. 3., 2004

[7] Manning, C. D., and Schutze, H.: Foundations of statistical natural language processing. *The MIT Press*. 1999.

[8] Jackson, P., and Moulinier, I.: Natural language processing for online applications: Text retrieval, extraction and categorization. *John Benjamins Publishing Co*. 2002.

[9] Bray, T., Paoli, J,, and Sperberg-McQueen, C., Extensible Markup Language (XML), *The World Wide Web Journal*, Volume 2, No. 4, pp 29—66, 1997.

[10] Lucene search engine: http://en.wikipedia.org/wiki/Lucene

[11] Baralis, E., Paraboschi, S., and Teniente, E.. Materialized views selection in a multidimensional database. In *VLDB Conference*, pages 156.165, 1997.

[12] Chaudhuri, S. and Dayal, U. An overview of data warehousing and OLAP technology. *SIGMOD Record*, Vol. 26, No. 1., pp 65--74, 1997.

[13] He, B., Wang, R., Chen, Y., Lelescu, A., and Rhodes, J.: BIwTL: A Business Information Warehouse Toolkit and Language for Warehousing Simplification and Automation. *Proceedings of the ACM SIGMOD*, Beijing, China, 2007.

[14] Brecher, J., Name=struct: A practical approach to the sorry state of real-life chemical nomenclature. Journal of Chemical Information and Computer Sciences, Vol. 39, No. 6, pp 943—950, 1999.

[15] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* Vol. 28, pp 31-36. 1998

[16] Leek, T.R., Information extraction using hidden markov models, Master's thesis, UC San Diego, 1997

[17] Lafferty, J., McCallum, A., Pereira, F., Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pp 282--289. Morgan Kaufmann, San Francisco, CA, 2001.

[18] Rodes, J., Boyer, S., Kreulen, J., Chen, Y., and Ordonez, P., Mining patents using molecular similarity search, In *Proceedings of Pacific Symposium on Biocomputing*, pp 304—315, 2007

[19] IBM Bluegene, http://www.research.ibm.com/bluegene.

[20] Spangler, W. S., Kreulen, J. T., and Newswanger, J. F.: Machines in the conversation: Detecting themes and trends in information communication streams. *IBM Systems Journal.* 2006.

[21] Hasan, M., Spangler, W. S., Griffin, T. D., and Alba, A, "COA: Finding Novel Patents through Text Analysis", *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France 2009, pp. 1175-1184., 2009

[22] Press, W. et. al.: Numerical Recipes in C. 2nd Edition. *New York: Cambridge University Press*. 1992. pp. 620-623.

[23] IBM Patents: http://www.ibm.com/ibm/licensing/patents/portfolio.shtml

[24] Cloud Computing: http://en.wikipedia.org/wiki/Cloud_computing