Combinatorial Document Matching: A Patent Case Study

Kas KASRAVI
Hewlett-Packard
West Bloomfield, MI 48323, USA

and

Kivanc OZONAT
Hewlett-Packard
Palo Alto, 94304 CA, USA

and

Claudio BARTOLINI
Hewlett-Packard
Palo Alto, 94304 CA, USA

ABSTRACT

We address a challenge in knowledge management and document processing that involves matching similar documents, especially when the matching is semantically one-to-many. Specifically, we propose a linguistic solution that automatically discovers a set of documents, which only in combination match a target document.

We apply the solution to patents, and in particular address the problem of detecting obviousness in patents. Detection of patent obviousness is generally a hard problem, since it involves finding a combination of relevant patents that, combined together, subsume the claims of a new patent application.

Our approach, based on combinatorial document matching, yields good results when applied to semantic analysis of the first independent claim of patents, therefore promising to save time and resources in patent prosecution, examination, and the discovery phase in patent litigation. Further, this approach lays a foundation for the broader problem of combinatorial document matching.

Keywords: Knowledge Management, Document Matching, Text Analysis, Linguistics, Patent.

1. INTRODUCTION

Due to the extensive amounts of digital documents available, it has become increasingly difficult for users to effectively sift through and examine such extensive document sets. In addition, document search, and particularly document matching, has been the subject of numerous research and commercial tools. Document matching is generally utilized for searching and clustering similar documents, organizing folders, and other content and knowledge management purposes.

Typically, a target document of interest is identified, and similar documents are linguistically matched against the target document on a one-to-one basis given their semantic similarity. In cases where the key concepts in a target document are present in combination within multiple documents, the user faces the tedious process of breaking down the concepts in the target document, performing partial matches, determining the relevance of the documents, and manually compiling a set of documents, which in combination, match the target document. We present an algorithm and a case study for automatically matching a set of documents that, in combination, are similar to a target document.

In this paper, our use case is patent search, and specifically performing obviousness detection, which is often a significant challenge in prior art search in patent prosecution and litigation cases. However, we believe the same approach can prove to be useful in other knowledge management use cases, such as gathering collateral for response to a Request for Proposal (RFP).

We describe a process for reliably detecting patent obviousness, where, “obviousness” is the legal standard for ensuring that inventions are not only new, but non-trivial and involve substantial ingenuity. Patent obviousness detection involves both qualitative and quantitative analysis. Our focus in this paper is quantitative analysis, in particular, flagging potential patents for obviousness purposes. This should save much time and resources in patent prosecution, examination, and the discovery phase of patent litigation.

Our work is currently focused on the semantic analysis of the first independent claim of U.S. patents.

2. PATENTS

Background

Patents embody a substantial amount of knowledge, and a driving force for economic growth [1]. Figure 1 demonstrates the correlation between patents and economic output by showing the correlation between the number of patents filed in U.S. states per 1,000 residents, and their Gross State Product (GSP), clustered into High ($48K/year), Medium ($39K/year), and Low ($34K/year) for year 2006.

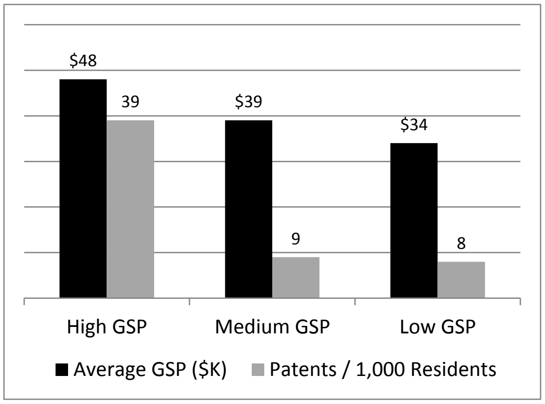

Figure 1. GSP and patent filing rates

The past two decades have seen a significant rise in patent activities (Figure 2) [2]. Along with that, the number of patent infringement actions grew at a compound average rate of 5.8% since 1991 [3].

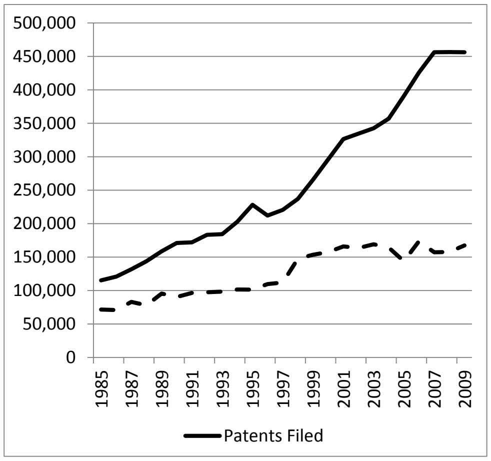

Figure 2. U.S. Patent Trends, 1985-2009

The processes for developing new inventions, prosecuting new patent applications, patent litigation, and other applications such as market intelligence, all require good insights into current and prior patents. The increasing level of patent activities is now creating a greater need for patent search and analysis.

Legal Foundation

The root of patents and their connection with economic growth are found in the U.S. Constitution, Article I, section 8, clause 8 authorizes the U.S. Congress "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries" [4]. The U.S. Congress has developed the U.S. patent law under Title 35 of the United States Code (35 USC) [5], which governs all patent practices. The key patentability requirements provided in 35 USC include:

- Subject Matter and Utility (35 USC 101)
- Novelty (35 USC 102)
- Non-Obviousness (35 USC 103)
- Specification (35 USC 112)

Our focus, the non-obviousness requirements, suggests that trivial inventions won't do. A patent requires an invention that demonstrates a substantial level of ingenuity, and difference from combinations of prior arts that would render the invention obvious in the mind of a person of ordinary skill in the art [6][8].

In its simplest form, quantitative obviousness examination is the identification of two or more prior arts, which in combination, teach the elements of a new invention. For example, if a new invention has elements ABXY, and a first prior art teaches elements AB, and a second prior art teaches XY, and then there is a possibility that the new invention is obvious in view of the first and second prior arts. In practice, most obviousness decisions are based on two prior arts, and they are usually (but not always) patents.

Therefore, the challenge that our solutions addresses is the identification of the key inventive elements in a new patent application, and examining whether those elements are taught in a combination of two or more patents. Our solution involves textual and semantic analysis of patent claims, focusing on the first Independent Claim, as described below.

Patent Claims

The Claims section is the most complex part of a patent document, and it also defines the legally protected invention, element by element. This section heavily uses legal language, and extracting meaningful concepts is challenging. However, we believe this is the most useful section of a patent for obviousness determination, as it unambiguously identifies all the inventive elements for each embodiment of an invention.

Patent claims represent knowledge in a very precise form and language. Some of the terms used in Claims are determined by court decisions, which in fact simplify the parsing and mining process due to the inherent structure and consistency. As the result, patent claims may be considered semi-structured data. There are several flavors of patent claims. The two most important types are Independent and Dependent Claims. Our work focuses on independent claims, which is briefly described below.

Independent Claims

An independent Claim is a single sentence that describes an independent embodiment of an invention in its broadest form, and does not refer to other claims. The general format of an Independent Claim is:

A <preamble> <transition> <body>.

Where,

- <preamble> introduces the invention;
- <transition> connects the preamble to the body via the use of the terms: “Comprising,” “Consisting,” and “Consisting essentially of,” where each phrase has a specific legal meaning.
- <body> is a recitation of the inventive elements and their relationships. An indefinite article (A or An) is used when an element is introduced the first time, and definite articles (The or Said) precede subsequent references.

An Independent claim typically begins with an indefinite article and ends with a period. The elements in the <body> are typically separated by a comma or semi-colon.

The following is a very simple example of an Independent Claim:

1. *A mousetrap comprising:*
   a) *a board,*
   b) *a spring attached to the board,*
   c) *a trigger attached to the spring, and*
   d) *a bait attached to the trigger.*

In this example, the invention is a mousetrap, and the inventive elements are the combination of a board, a spring, a trigger, and bait.

3. CHALLENGE OF ELECTRONIC PATENT SEARCH

Electronic patent tools have been available for some time. For example, Google Patents [9], Free Patents Online [10], United States Patent and Trademark Office (USPTO) [7], and others provide patent databases and search tools, which help to identify prior arts. These tools primarily rely on searching by the structured fields (e.g., dates, inventors, and assignees), as well as keywords.

Patent search is a difficult knowledge management challenge, due to variations in linguistic expressions, inconsistencies in classification, different legal interpretations, and different styles of patent drafting.

Searching for obviousness adds the additional challenge of partial search; where, only in combination, multiple documents identify the prior art.

Therefore, the primary challenge that our work addresses is the discovery a set of documents, which only in combination, match a target document based on a certain set of concepts.

4. ALGORITHM AND CASE STUDY

We developed an algorithm to examine the obviousness of a target patent application, via analyzing a repository of prior art patents. To test our algorithm, we used U.S. patent application 20030120517, which was rejected due to obviousness in view of patents 6092039, 6754631, 6651060, and patent application 20020194476. The target patent belongs to the patent class 705/3, which is “patient records management” in the general class of “data processing” inventions. Since this patent application is an IT patent, we consider IT to be our analogous art, which is generally included in patent classes 700-726. Patents in all other classes are considered to be non-analogous art (non-IT) for the purpose of our experiment. Our goal was to assess whether our algorithm could successfully detect the prior art patent documents from a large set of analogous and non-analogous patents.

Repository

Our target patent application is an invention that transmits, converts to text, and prints the dialog data between a physician and a patient. The first independent claim of the patent is as follows:

A dialog data recording method in a system in which a computer terminal of a person, capable of entering audio data and an image and of data printing, is capable of communicating with a server managed by a third person and having a large-capacity memory apparatus, the method comprising steps of: transmitting dialog data, containing an image and audio data in a dialog between said person and another person, from the computer terminal of said person to said server for storage in said large-capacity memory apparatus; recognizing and converting the audio data in said dialog data into text data; generating script data based on said text data; and printing said script data by the computer terminal of said person.

The claim of the invention above was challenged by the USPTO on the issue of non-obviousness. In particular, the USPTO stated that the claim was obvious given the two claims from the two previous patents. The first of these claims was:

A method for automatic speech recognition (ASR) and vocoding (VC), comprising the steps of: converting a first signal representing speech to a second signal having raw mel capstrum vector (MCV) and a third signal having raw pitch; subtracting a calibration vector from said MCV to form a difference vector; multiplying a calibration matrix with said difference vector to produce a recalibrated MCV; recalibrating said raw pitch with a logarithmic function; concatenating said recalibrated MCV with said recalibrated pitch to form a recalibrated vector; compressing and quantizing said recalibrated vector to form a vector quantized signal; and forwarding said vector quantized signal to a remote receiver for decoding said vector quantized signal received by the remote receiver to recover said speech.

And, the second one was:

A method for memorializing a conversation of a plurality of speakers, comprising: sampling a sample utterance of each of the plurality of speakers thereby producing a plurality of sample utterances each having a characteristic corresponding to a corresponding one of the speakers; associating a characteristic of the sample utterances with the corresponding one of the plurality of speakers; recording the conversation by saving the conversation to a storage medium as the conversation occurs whereby conversation utterances generated by each of the plurality of speakers during the conversation are saved to the storage medium; identifying the one of the plurality of speakers who generated each one of the conversation utterances by matching the characteristic of the sample utterance associated with the corresponding one of the plurality of speakers with a characteristic of the conversation utterance; associating information regarding the identified speaker with at least one of the conversation utterances; generating a transcript of the conversation, the transcript including the information associated with the at least one of the conversation utterances; and interpreting the transcript of the conversation and generating a summary of the conversation based upon said interpreting the transcript.

We randomly selected 500 IT-related patents from a repository of 5,000 IT patents with USPTO classification code ranging between 700 and 726. We also randomly selected 500 non-IT related patents from a repository of 5,000 non-IT patents. Thus, our test set includes 1,000 patents, half of which are IT-related patents and the remaining half are non-IT related patents.

We also included in our test set the patent documents known to be obviating the target patent.

Algorithm

We devised an algorithm that compares each pairwise combination of claims in our test set to the target claim. If the target claim has a high similarity with the compared pairwise test claims, the pair is ranked highly.

Our algorithm consists of the following four components:

Claim parts: We break down each claim into its inventive elements at the semi-colons. If semi-colons are not present, we use commas. This procedure follows from the general structure of a claim as explained in section 4.

Keyword Extraction: For each element of each claim in our test set, we extract the keywords and key phrases of the elements.

Relevance: Given two keyword sets, the target claim keyword set and the test pair claim keyword set, we compute the similarity between the two sets of claim keywords.

Ranking: We rank the pairwise combinations of claims in our test, based on how similar they are to the target claim.

For the keyword extraction, we used the online keyword extraction tool provided by Yahoo [11]. The tool accepts a paragraph (in our case, the claim) as input, and outputs a set of keywords and key phrases. Given a target claim and a pair of test claims, we denote the keywords of the target by P, and the keywords of the test pair by S. S consists of N subsets of keywords for each of its N elements, and P consists of M subsets of keywords for each of its M elements.

Given a set S of keywords and key phrases for a test claim, and the P of keywords for the target claim, we estimate the similarity between S and P. In a given repository of documents, the existence of many documents that contain both the keywords in S and the keywords in P indicates that the sets S and P are likely to be relevant. We use the Web as the document repository, and use the Yahoo search engine results as a proxy to estimate the number of documents common to both P and S.

We denote by A any subset of P, and by B any subset of S. We record |A|, the number of documents that Yahoo retrieves in response to A; |B|, the number of documents that Yahoo retrieves in response to B; and, |A,B| the number of documents that Yahoo retrieves in response to A and B. The similarity between A and B is computed as min(|A|,|B|)/|A,B|. Given any A, the subset B of S that maximizes the similarity ratio is taken as A's counterpart in S.

Given P and S, their similarity is taken as the sum of the similarity ratios of the counterpart subsets of (A's and B's) of P and S.

Our algorithm can be viewed as a soft-version of counting the number of elements in the target claim similar to at least one element in either of the test pair claims. We replace the count by a similarity ratio.

Individual Rankings
We used our algorithm to rank the individual patents in our repository based on how similar each is to the target patent.

We conducted two experiments to test the individual ranking of the patents. In the first experiment, we ranked the 500 IT patents and the four known prior art patents based on their similarity to the target patent. The rankings of the prior art patents are shown in the first column of Table 2 (column label IT-I). Two of the four prior art patents were ranked in the top 10, one ranked in the top 20, and the other ranked in top 40.

In our second experiment, we ranked the 500 non-IT patents and the four prior art patents based on their similarity to the target patent. The rankings of the prior art patents are shown in the third column of Table 1 (column label N-I). Three of the four prior art patents were ranked in the top 10, and one ranked in the top 20. Two of the prior art were ranked in the top five patents.

In Table 2, we show the top five non-IT patents in our ranking (column label I), and in Table 3, we include the top five IT patents (column label I). We note that the top patents in both cases include keywords related to image engineering. The top non-IT patent, for instance, has the keyword "touch screen," while the next non-IT patent in the ranking has the words "front panel" and "back panel." We notice a similar trend in Table 3. The top two patents include keywords such as "image representative," "composite image," "images," and "screen."

The observations above from Tables 2 and Tables 3 are not surprising since the target patent contains the keywords "data printing" and "image printing." This can be viewed as an indication that our approach of using the World Wide Web to compute relevance of patents is reasonable.

Pairwise Rankings
We also compared the pairwise rankings of the patents based on their similarity of the target patent. To have a fair comparison among the three sets of patents (prior art of Table 1, IT of Table 2 and non-IT of Table 3), we paired the prior art patents with other prior art patents, the IT patents with other IT patents, and the non-IT patents with other non-IT patents. We did not pair, for instance, IT patents with non-IT or prior art patents.

Our results indicated that the pairs that were ranked at the top are the patents that ranked high individually. This is expected, since the set of keywords of a pair is the union of the keywords of the individual patents in that pair.

A more significant conclusion drawn from the pairwise rankings, however, is that pairwise ranking improves the ranking of the prior art patents, while the rank of the remaining (both IT and non-IT) patents are reduced.

The results in Table 1 confirm the above conclusion. From Table 1, we note that all four prior art patents are ranked in the top 15, with one of them ranking in the top five, when tested with IT patents (column label IT-P). Table 1 also shows that all four prior art patents are ranked in the top 10 when tested with non-IT patents (column label N-P).

5. DISCLAIMERS

1- This paper references a number of topics related to patent law; however, this paper does not provide legal advice in any form or fashion. The reader is encouraged to contact a patent attorney or appropriate legal counsel for legal advice.

2- The analysis presented in this paper uses publicly available data from the United States Patent and Trademark Office (USPTO) [7]. The analyses presented in this paper are for scientific illustration purposes only, limited in scope, and not intended to express an opinion about any of the patents or their inventors and/or assignees.

6. CONCLUSION

Patents represent a substantial body of knowledge, and can be of critical value to forward looking organizations. Publicly available patent databases contain a substantial amount of patent data, and the increase in patent activities are demanding better methods for detecting obviousness of new patent applications.

To the best of our knowledge, no good methods exist to detect obviousness of new patent applications, resulting in time consuming and resource intensive effort during patent prosecution, examination, and the discovery phase of patent litigation.

Through a series of experiments on publicly available data from the USPTO database [7], we showed that our approach, based on combinatorial document matching applied to semantic analysis of the first independent claim of patents yields very good results.

8. REFERENCES

[1] Jarboe, K.P., and Atkinson, R.D., **The Case for Technology in the Knowledge Economy: R&D, Economic Growth, and the Role of Government**, Progressive Policy Institute, June 1998.

[2] **World Intellectual Property Indicators 2010**, http://www.wipo.int/ipstats/en/statistics/patents/

[3] Levko, A., Torres, V., and Teelucksingh, J., **2008 Patent Litigation Study: Damagaes, Awards, Success Rates and Time-to-Trial**, PriceWaterhouseCoopers LLP, 2008.

[4] **United States Constitution** – Article I.

[5] http://uscode.house.gov/download/title_35.shtml

[6] Graham et al. v. John Deere Co. of Kansas City et al., 383 U.S. 1 (1966).

[7] http://www.uspto.gov

[8] KSR Int’l Co. v. Teleflex, Inc., 550 U.S. 398 (007).

[9] http://www.google.com/patents

[10] www.freepatentsonline.com

[11] http://developer.yahoo.com/search/content/V1/termExtraction.html

Table 1. Individual (denoted by I) and Pairwise (denoted by P) Rankings for the Prior Art Claim

Rank (IT-I)	Rank (IT-P)	Rank (N-I)	Rank (N-P)	First Independent Claim (excerpts)	Sample Keywords
8	5	2	1	A method for automatic speech recognition (ASR) and vocoding (VC), comprising the steps of: converting a first signal representing speech to a second signal having raw mel capstrum vector (MCV) and a third signal having raw pitch; subtracting a calibration vector from said MCV to form a difference vector; …	automatic speech recognition, difference vector, vector c, logarithmic function, mcv, concatenating, asr, calibration, pitch, matrix
10	11	3	2	An apparatus having a digital protection mechanism, comprising: a tangible object; a digital protection system attached to said tangible object, said digital protection system comprising: (a) an external interface for receiving data requests; (b) a processor coupled to said external interface, said processor capable of transforming data …	internal data storage, encryption algorithm, first transformation, private key encryption, external interface, attribute data, tangible object, digital signature, protection mechanism, data requests, public key
18	14	8	5	A method for memorializing a conversation of a plurality of speakers, comprising: sampling a sample utterance of each of the plurality of speakers thereby producing a plurality of sample utterances each having a characteristic corresponding to a corresponding one of the speakers; associating a characteristic of the sample utterances with the corresponding one of the plurality of speakers; …	plurality, storage medium, utterances, utterance, sampling, speakers
35	15	18	10	In a data processing center, included in a system that also includes one or more requestors of records and a plurality of providers that generate or have custody of records, a method for processing requests for records comprising: receiving from a requestor a request for a record in the custody of a provider; …	data processing center, requestor, authorizations, plurality, relationship

Table 2. Individual (denoted by I) and Pairwise (denoted by P) Rankings for the Non-IT Patent Claims

Rank (I)	Rank (P)	First Independent Claim (excerpts)	Sample Keywords
1	3	A method of operating a control console comprising: displaying information on a console display; detecting the presence of a docking module on the console display; and changing the information presented on the console display in response to said step of detecting to convey information tailored specifically to user operation of the docking module; wherein the docking module is a special purpose input module, and the information tailored specifically to user operation of the special purpose input module comprises legends that are displayed under corresponding legend areas of the special purpose input module when the docking module is docked and overlays the display; …	push button switches, input module, independent position, special purpose, plurality, input devices, touch screen, segments, segment, legends, array, presence,
4	4	An enclosure assembly comprising: an enclosure including a back panel, first and second side walls extending from said back panel, a top end, a bottom end, and a front panel; an electrical switching apparatus housed by said enclosure and including a handle; and an operating mechanism coupled to said handle of said electrical switching apparatus, said operating mechanism comprising: …	side walls, linkage, back panel, front panel,
5	6	An angle mounting bracket and structural bridging fastener comprising: an angle mounting bracket that is attachable to a stable base, the bracket comprising a substantially planar main body and a mounting plate oriented perpendicular to the main body; said mounting plate and main body being connected through a complex formation in the angle mounting bracket; …	subcomponent, attachment point, stable base, mounting bracket, fastener, mounting plate

Table 3. Individual (denoted by I) and Pairwise (denoted by P) Rankings for the IT Claims

Rank (I)	Rank (P)	First Independent Claim (excerpts)	Sample Keywords
1	1	A method for on-line viewing of articles, comprising: providing a host-site that is web accessible to an on-line viewer and web-linkable to different article-provider sites, the article provider-sites having images of articles for view via the web; linking the on-line viewer to the host-site and receiving a command from the on-line viewer that selects a structure and at least one type of the articles; using the host-site in response to the command, …	image representative, virtual closet, host site, composite image, web link, images
2	2	A method of modifying a display order of user interface (UI screens, comprising the steps of: providing a single record text based setup data file for a suite installation and setup application having at least one section containing a display order textual listing of the user interface (UI) screens; providing a text editor; ...	setup application, user interface, screens, setup data,
3	3	A computer-implemented method for diagnosing a problem in a product using a Bayesian super model data structure which stores a predetermined set of problems, predetermined criteria for identifying problems in the set, and sub model data problems, predetermined criteria for identifying problems in the set, and sub model data structures including actions for addressing the problems in the set, the method comprising: receiving user input including criteria for identifying the problem; …	model data, data structure, super model, data structures, probability, execution, match
4	4	A method for recording data onto an optical disc, comprising: generating a set of pointers to associate record data structures with a writing order, the set of pointers defining a dynamically ordered list of record data structures; processing each of the record data structures one after another in the writing order to produce an ordering data structure for each file in a set of files to be recorded onto the optical disc, …	optical disc, data structure, data structures, pointers