Innovation Analytics Using Mined Semantic Analysis

Walid Shalaby, Wlodek Zadrozny
Computer Science Department
University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223, USA
{wshalaby, wzadrozn}@uncc.edu

Abstract

In this paper we describe our work on cognitive assistance (Cog) technology in the innovation analytics domain. We propose a framework for innovation analytics and management using Mined Semantic Analysis (MSA). Our goal is to build a semantic driven visual interactive analytics engine that provides insights on innovation data using conceptual knowledge derived from huge unstructured textual knowledge corpora (e.g., Wikipedia). Throughout the paper we demonstrate a case study utilizing our framework for providing computational assists on competitive intelligence by automatically defining the innovation portfolio of an organization, and using that information to identify other key players with similar portfolios which could be candidates for acquisition.

Introduction

Patents and innovations represent proxies for economic, technological, and even social activities. Therefore, patent analysis has received considerable attention in the literature\(^1\) (Far et al. 2015; Zhang et al. 2015; Shalaby and Zadrozny 2015b; Cormack and Grossman 2014; Mahdabi et al. 2013; Lupu et al. 2011; Koch et al. 2011).

Typical innovation management use cases include: 1) Technology exploration in order to capture new and trendy technologies in a specific domain and subsequently using them to create ideas for new innovative services, 2) Technology landscape analysis in order to assess the density of patent filings of specific technology and subsequently direct R&D activities accordingly, 3) Competitive analysis and benchmarking in order to identify strengths and differences of corporation’s own patent portfolio compared to other key players working on related technologies, 4) Patent ranking and scoring in order to quantify the strength of the claims of an existing or a new patent, and 5) Prior art search in order to retrieve patent documents and other scientific publications relevant to a new patent application. All those innovation management activities require tremendous level of domain expertise which, even if available, must be integrated with highly sophisticated and intelligent analytics that provide cognitive and interactive assistance to the users.

Due its technical nature, patents language tends to be highly sophisticated with complex vocabulary, jargon, and domain specific terminology. Despite those linguistics challenges, most research in automated patent analysis is inspired by either content-based (e.g., term co-occurrence) or metadata-based (e.g., bibliographic data) methods.

We, alternatively, embrace semantics driven analysis of innovation data. Our hypothesis is that, by subtle incorporation of external conceptual knowledge, we could bridge the linguistic and domain expertise gaps and provide non-expert users cognitive assistance that would not be achievable by using the limited content-based approaches.

To this end, we propose a semantic framework for innovation analytics. We utilize Mined Semantic Analysis (MSA) (Shalaby and Zadrozny 2015a), a novel distributional semantics approach which employs data mining techniques. MSA constructs a conceptual knowledge graph whose nodes are encyclopedic concepts and links are quantified associations between those concepts. MSA builds that knowledge graph offline by mining for latent concept-concept associations in a target encyclopedic textual corpora (e.g., Wikipedia) using association rules mining. Once constructed, this rich knowledge graph can be used for several tasks like semantic search, concept expansion, measuring semantic relatedness, word sense disambiguation, resolving vocabulary mismatch, and others.

Mined Semantic Analysis

Our innovation management framework utilizes MSA, a novel distributional semantics technique which proved effectiveness for evaluating semantic similarity/relatedness on benchmark data sets (Shalaby and Zadrozny 2015a). MSA stands unique from other explicit semantic analysis approaches like ESA (Gabrilovich and Markovitch 2007) and SSA (Hassan and Mihalcea 2011) as it maps textual content into a conceptual space that captures not only explicit keyword-concept associations but also latent concept-concept associations.

MSA builds two repositories in order to map seed keyword/text to the conceptual space: 1) A search index of all documents in a target encyclopedia (e.g., Wikipedia), and 2) A knowledge base of latent concept-concept associations learned using association rules mining (Agrawal, Imieliński, and Swami 1993) of Wikipedia “See also” link graph.
index is used to construct an initial set of concepts (articles) explicitly mentioning the seed text (explicit concepts). For each explicit concept, MSA retrieves its associated concepts from the association rules knowledge base (latent concepts). Both explicit and latent concepts represent the conceptual mappings of the seed text.

Figure 1 shows the conceptual representation for the Abstract section of this paper. We show the top 4 explicit concepts (light blue nodes) along with the top 8 latent concepts associated with each of them (red nodes). It is important to mention that, explicit concepts are ranked top-down according to their relevance to the seed text. Latent concepts of each explicit concept are also ranked top-down according to their relevance to the explicit concept based on the support of the association rule containing both of them.

As we can see in Figure 1, MSA could identify semantically related concepts to the Abstract section including Text mining, Big data, Strategic management, and Innovation. The latent concepts serve as a powerful mechanism for concept expansion. They augment the knowledge required to capture key ideas expressed in the seed text in multiple ways offering hypernymy/abstraction (Strategic Management and Management), hyponymy/specificity (Text mining and Text classification), synonymy (Innovation and Invention), and relatedness/associativity (Big data and Internet of Things).

Case Study

In order to demonstrate the viability of our semantic driven framework in the innovation analytics domain, we present a case study on competitive intelligence. The case study explains how MSA can be applied to: 1) define the Intellectual Property (IP) portfolio of an organization, and 2) identify other key players with similar IP portfolios which could be candidates for acquisition.

To define the IP portfolio of an organization, we built a big index of all US granted patents\(^2\) between 1976 and Oct. 2014. We used Apache Solr to build and search the index. The total index size was about 200GB comprising around 4.7 million documents. For each patent, we indexed its title, abstract, description, claims, assignee, and publication date.

The scenario starts with a seed organization and ends with potential key players with similar IP portfolios. In the process, target organization’s IP portfolio is defined in terms of technological and technical concepts expressed explicitly or implicitly in the organization’s patents.

We exemplify by considering Bank of America\(^3\) (BoFA) as a target organization. By searching our patents index, we found approximately 790 patents whose assignee is BoFA. IP portfolio identification is a multi-step process that leverages representative description of the patents, e.g. their titles, abstracts, descriptions, and/or claims. We extract titles of 100 patents at random (see Table 1 for sample titles) as a representative description of BoFA innovations. Then, we pass all titles as a single snippet to MSA to discover the corresponding IP concept space. Figure 2 shows MSA’s top 20 relevant concepts which represent BoFA’s IP portfolio using

\(^2\)https://data.uspto.gov/uspto.html
\(^3\)https://www.bankofamerica.com/
Figure 2: Concept graph using BofA’s 100 patent titles. Light blue nodes are explicit concepts and red nodes are latent ones.

Figure 3: Concept graph using Witricity’s 10 patent titles. Light blue nodes are explicit concepts and red nodes are latent ones.
titles from Table 1. As we can notice, these concepts are semantically related to titles in Table 1.

The final step in our competitive analysis scenario is to identify key players with similar IP portfolios to BofA. To do this step, we take the top ranked concepts and combine them to construct a search query against the patents index. We limit the search to patent claims as they define in technical terms the scope of protection sought by the inventor. Among the top ranked key players in the competitors list, we can find companies like ActivIdentity\textsuperscript{4}, which is specialized in identity assurance, SecureEnvoy\textsuperscript{5} which is specialized in authentication and verification, and IBM\textsuperscript{6}.

To validate the robustness of our semantic framework, we repeated the same competitive analysis experiment on Witricity\textsuperscript{7}, which is specialized in wireless energy transfer using resonant magnetic coupling. Table 2 shows titles of 10 Witricity’s patents. Figure 3 shows the top 20 relevant concepts representing Witricity’s IP portfolio using those titles. We validated the relevance of retrieved concepts to the wireless energy industry based on feedback from a domain expert. To close the loop, we retrieved the list of similar key players that included Qualcomm\textsuperscript{8}, Powermat\textsuperscript{9}, and Mojo Mobility\textsuperscript{10} which all provide wireless charging solutions.

**Conclusion**

We presented our ongoing research on cognitive assistance (Cog) technology in the innovation analytics domain. Through the paper, we demonstrated a case study using MSA’s semantic driven framework for competitive intelligence. Future work includes extending this research through effective visualization, interaction, and knowledge incorporation for boosting human intellectual effectiveness.

**References**


