Expert Systems With Applications 42 (2015) 9389–9401

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A framework to explore innovation at SAP through bibliometric analysis of patent applications

Tabitha L. James [a,*], Deborah F. Cook [a], Sumali Conlon [b], Kellie B. Keeling [c], Stephane Collignon [a], Trevor White [a]

[a] Virginia Tech, 1007 Pamplin Hall, Blacksburg, VA 24061, United States
[b] University of Mississippi, 247 Holman Hall, University, MS 38677, United States
[c] University of Denver, 2101 S University Blvd., Denver, CO 80208-8931, United States

ARTICLE INFO

Keywords:
Software
Innovation
Enterprise data management
Business intelligence and social media/data analytics
Text analysis
Cluster analysis

ABSTRACT

Easily accessible patent databases and advances in technology have enabled the exploration of organizational innovation through the analysis of patent records. However, the textual content of patents presents obstacles to gleaning useful information. In this study, we develop an expert system framework that utilizes text and data mining procedures for analyzing innovation through textual patent data. Specifically, we use patent titles representing the innovation activity at one company (SAP) and perform a bibliometric analysis using our proposed framework. Enterprise software, of which SAP is a pioneering developer, must serve a wide assortment of functions for companies in many different industries. In addition, SAP's sole focus is on enterprise software and it is a market leader in the category with substantial patent activity over the last decade. Using our framework to analyze SAP's patent activity provides a demonstration of how our bibliometric analysis can summarize and identify trends in innovation in a large software company. Our results illustrate that SAP has a breadth of innovative activity spread over the three-tier software engineering architecture and a lack of topical repetition indicative of limited depth. SAP's innovation is also seen to emphasize data management and quickly integrate emerging technologies. Results of an analysis on any company following our framework could be used for a variety of purposes, including: to examine the scope and scale of innovation of an organization, to examine the influence of technological trends on businesses, or to gain insight into corporate strategy that could be used to aid planning, investment, and purchasing decisions.

1. Introduction

Innovation is of interest to academics and practitioners alike. It has been examined for decades using several approaches and for a variety of means. Innovation's link to company performance has been extensively studied (Leidner, Lo, & Preston, 2011; Noruzy, Dalfard, Azhdari, Nazari-Shirkouhi, & Rezazadeh, 2013; Srivardhana & Pawlowski, 2007). Research has approached innovation at multiple levels. At the most coarse-grained levels, the influence of innovation on national or global economies is of interest (Hicks, Breitzman, Olivastro, & Hamilton, 2001; Slater, 2012). In many cases, innovation across an industry has been of primary interest (Pullen, de Weerd-Nederhof, Groen, & Fisscher, 2012; Tajeddini & Trueman, 2012). These studies explore innovation as a competitive advantage (Adegoke, Walumbwa, & Myers, 2012), to aid in evaluation of merger activity within an industry (Breitzman & Mogee, 2002; Breitzman & Thomas, 2002), as an assessment of service quality (Song, Song, & Di Benedetto, 2009), to explore emerging technologies within an area (Daim, Rueda, Martin, & Gerdsri, 2006; Kim, Suh, & Park, 2008), as a tool for sustainability/green initiatives (Bos-Brouwers, 2010; Chen & Chang, 2013), or to examine technological strategies within an industry (Ashurst, Freer, Ekdahl, & Gibbons, 2012; Han, Kim, & Srivastava, 1998; Schmoch & Schnoring, 1994). At the organizational or firm level, management of innovation (Sundbo, 1997) and enhancement of business planning (Lee, Yoon, Lee, & Park, 2009; Zippel-Schultz & Schultz, 2011) may be the goal.

Various researchers have approached the study of innovation using patent data. Patent data has been examined using a variety of approaches and procedures for several purposes. Hicks et al. (2001) used patent bibliometrics to investigate the changing composition of innovative activity in the US. They noted patent indicators pointing to significant innovation in information and health technologies. Daim et al. (2006) point to the difficulties of forecasting when there is no historical data available and used a combination of bibliometrics and

* Corresponding author. Tel.: +1 540 231 3163; fax: +1 540 231 3752.
E-mail addresses: tajames@vt.edu (T.L. James), dcook@vt.edu (D.F. Cook), sconlon@bus.olemiss.edu (S. Conlon), kellie.keeling@du.edu (K.B. Keeling), stephane@vt.edu (S. Collignon), tsw90@vt.edu (T. White).

http://dx.doi.org/10.1016/j.eswa.2015.08.007
0957-4174/

patent analysis to forecast emerging technologies in the areas of fuel cells, food safety, and optical storage.

Siwczyk, Warschat, and Spath (2012) note that patents are used to protect innovative ideas and describe the use of patents as a source of new ideas in the early phases of technology management processes. They also discuss the need for software solutions to facilitate patent analysis as patent databases continue to grow. To advance software to explore patents, Kim et al. (2008) used a k-Means clustering algorithm to create patent maps for ubiquitous computing technology. Shih, Liu, and Hsu (2010) propose a patent trend change mining approach to examining patent activity and use it to examine trends in industry and company activities in Taiwan's semiconductor industry. Lee, Yoon, and Park (2009) used a combination of text mining and principal component analysis to identify "patent vacancies" in patent maps to help identify new technology creation opportunities. Yoon and Kim (2011) noted that patents are good sources of information related to technological change while Yoon, Park, and Kim (2013) used natural language processing of patent text to identify technological competition trends finding "patent vacuums" and "technological hot spots". Chen (2009) used experts to assist in the formation of a patent map for car industry patents. Chen (2009) suggests that patent maps can be created using the process detailed in the study by any company that wishes to use the resulting patent map as a strategic tool.

Previous research indicates that analyzing patent data can lead to many beneficial outcomes, such as detecting trends in innovation and forecasting emerging technologies. It has also been suggested in previous research that methodologies for conducting patent analysis, especially textual content, may be useful to companies in identifying strategic opportunities in a company's own patent activity, comparing innovation or R&D approaches between units, companies or industries, or identifying industry opportunities. Patent analysis as a means to assess innovation could also be leveraged in merger and acquisition decisions or large-scale capital purchases where the health and innovation status of the company is important in assessing access to future product development and support.

In the current study, we propose an expert system framework for analyzing innovation through patent data that combines techniques from text and data mining to conduct temporal and textual analyses of patent activity. We utilize a combination of text and data mining procedures similar in nature to many of the procedures seen in the previous expert system approaches to patent analysis described above. However, our approach combines different procedures and stages of analysis than previous studies, that depending on the desired outcome, could provide a more robust and flexible analysis. Our framework combines traditional text and data mining techniques and a combination of open source, freeware, commercial, and research tools is used to implement the procedures. Our framework is flexible enough that any similar tool or relevant procedure could be substituted or added respectively.

To illustrate our framework, we perform a bibliometric analysis of SAP's patent data. SAP is the market leader for enterprise resource planning software (more generically called enterprise software). Enterprise software is by definition complex and wide-ranging. To adequately handle the system infrastructure needs of large companies or organizations, the software must contain logic for all business functions. Enterprise software serves as the information technology backbone for a company, and as such, needs to be customizable and flexible to meet the needs of organizations that vary in size, scope, purpose, etc. Competitors of SAP range from smaller software companies focused on a particular business function (e.g., HighJump Software and SalesForce) to large, diversified software companies that offer enterprise software as part of a larger portfolio of other software products (e.g., Oracle and Microsoft).

Software is a dynamic and innovative subset of the technology industry and SAP provides one of the most complex examples of software that is heavily utilized by business. In addition, SAP's sole focus is to produce and support enterprise system software. SAP is reported as the market leader in the business-management software market with a 24% market share according to Gartner, Inc. (Norton, 2014). Oracle is reported as the next closest competitor with a 12% market share (Norton, 2014). Sage, Infor, and Microsoft are reported to make up a combined 17% of the market with a large number of companies holding market shares of 3% or less (http://www.statista.com/statistics/249637/erp-software-market-share-by-company/) making up the remaining market share. Nearly 80% of the Fortune 500 companies use SAP software and 63% of the financial transactions are at least partially processed with SAP software (http://fortune.com/2012/03/29/inside-saps-radical-makeover/).

SAP has a history of innovation (Leimbach, 2008), currently demonstrated by SAP Predictive Analytics in addressing the ongoing challenges of big data and SAP HANA (High-Performance Analytics Appliance) in the area of the in-memory computing needed to facilitate predictive analytics. It has been noted that SAP's growth has been assisted by its focus on in-memory computing driven by HANA that improves performance by attempting to minimize time spent transferring data from permanent storage, in part because SAP was one of the first to move into this area (Ashford, 2011). Of all other enterprise software companies SAP has the most patent activity and is the only company with substantial patent activity that is focused solely on enterprise software, making SAP the ideal option to study innovation within a single software developer in this industry. Thus, SAP provides a good case study for our framework that may shed insight into the innovation strategies of one of the world's most important corporate software providers.

The rest of the paper is organized as follows. In the next section, we review the history and development of enterprise systems to show the historical evolution of this software category. This provides a historical presentation of a quickly evolving industry and illustrates the pace of technological and business function integration. The review motivates our choice of enterprise software as a case study and assists with the interpretation of the findings from the application of the framework. In Section 3, we will describe the data collection and methodology, following with the analysis and discussion in Section 4. We offer conclusions and directions for future work in Section 5.

2. Background and motivation for case study

2.1. Enterprise systems and SAP

Since the 1960s, computer systems for enterprise management have grown, now encompassing all aspects of an organization, and have branched out into activities related to supply chain management. The development of software programs to help companies manage specific processes (e.g. inventory and order management) presented attractive possibilities as computer use in business proliferated throughout the 1960s. The process of integrating all activities of a firm into a single application eventually evolved into "enterprise systems" (ES) that attempt to allow firms to electronically manage most (or all) of their business activity (Collignon, James, & Cook, 2010). Collignon et al. (2010) developed a time line to illustrate the development of Enterprise Systems from their beginnings as inventory management systems, through their evolution into Material Requirements Planning and then to Enterprise Resource Planning (ERP) software and beyond (Fig. 1).

The term ERP was introduced early in the 1990s to differentiate the fully integrated business applications software of SAP and others from the business software of older companies (Campbell-Kelly, 2003). ERP software systems provide the information backbone to support the infrastructure of many companies. Standard functions such as ordering materials, scheduling production, sales and operations planning, and general accounting functions are managed by

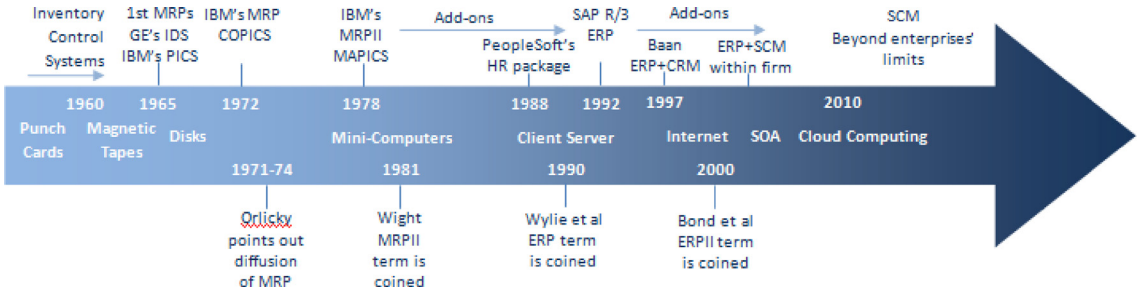

Fig. 1. ERP timeline.

Table 1
SAP solution categories (http://go.sap.com/solution.html).

Analytics	Content and collaboration
Enterprise resource planning	Financial management
Supply chain management	Supplier relationship management
Customer relationship management	Human capital management
Data management	

ERP systems and initial implementations of ERP systems focused on these types of transactional and record keeping tasks within enterprises (Holtsapple & Sena, 2005). Billions of dollars have been spent by enterprises to buy and implement ERP systems and the purchase and implementation of ERP software by any company is a major endeavor.

Enterprises regularly use the information provided by ERP systems to facilitate improved integrated decision-making based on system level conditions and considerations. Palaniswamy and Frank (2000) describe the need for organizations to work quickly with other organizations to make strategic decisions. ERP systems provide a mechanism to facilitate this type of strategic action across enterprises. Yusuf, Gunasekaran, and Abthorpe (2004) and Yao and He (2000) report that ERP capabilities are increasingly important because of the ability of ERP to integrate the flow of financial and operations informations as well as material and resources to support strategic activities. ERP systems are the infrastructure to provide the data and information needed for companies to analyze innovation possibilities for new product offerings and new market entrances, thus supporting the continuing evolution of agile, demand-driven supply chains.

SAP, the world's largest provider of business software (Campbell-Kelly, 2003; Wolde, 2012) including the best known of these enterprise management systems, was developed by SAP AG in Germany in the early 1970s. SAP is the market share leader with 24% of the world market according to Gartner, Inc. (Norton, 2014). The dependence of major companies throughout the world on SAP software is significant as noted by Campbell-Kelly (2003), who stated that if SAP version R/3 were to cease to exist: "the industrial economy of the Western world would come to a halt, and it would take years for substitutes to close the breach in the networked economy".

SAP provides a number of solution categories (Table 1) and reports that they have more than 291,000 customers in 190 countries (http://www.sap.com/corporate-en/our-company/index.epx). Additionally SAP provides industry-specific solutions for more than 25 industries (Table 2). The SAP ERP system remains the most deployed of the SAP solutions (Missbach, Stelzel, Gardiner, Anderson, & Tempes, 2013).

SAP's 2014 revenues were reported to be €17.56 billion, which is a 4% increase from 2013 (Forbes, 2015). SAP is number 190 on the Forbes' world's biggest public companies 2015 ranking with $23.29 billion in sales and 74,406 employees.[1] The annual growth rate of SAP revenues between 1973 and 2006 ranged from −5.2% (the only negative rate) to 104.8% in 1989 (Leimbach, 2008). The average revenue growth rate during that 34-year time period was 37.5%. Appendix A provides a brief history of SAP.

3. Data collection and methodology

3.1. Data collection

The U.S. Patents and Trademark Office (USPTO) provides a searchable database of patent applications from the years 2001–present. Patent applications were used rather than issued patents due to the fact that it typically takes 4–6 years for a patent to be issued, which makes issued patent data less complete. Given the long delay in appearance of issued patents and the number of other factors that have a role in the final acceptance of a patent, patent applications provide a more complete and robust view of the innovative activity of SAP. However, while less pronounced, there is still a lag in the recording of patent applications. Therefore, at the time of collection, years 2011 and 2012 may still not be fully complete. It can be seen in Table 3, that the number of patent applications obtained from these two years is still quite large in comparison to SAP's yearly activity prior to 2011. In addition, while we collected what had been filed thus far in 2013, the date of collection was February 13th. Thus, 2013 is only a partial year collection.

We searched the USPTO database (http://patft.uspto.gov/) for patent applications where the assignee name was SAP. Following the database convention, the query string used was "an/SAP". The "an" in the query string refers to assignee name, which by the USPTO definition is: "the name of the individual or entity to whom ownership of the published application was assigned at the time of publication" (http://appft.uspto.gov/netahtml/PTO/help/helpflds.html). As of the collection date, SAP was the assignee on 1724 patent applications from 2001–February 2013.

3.2. Methodology

To explore SAP's innovative activity, we utilized text and data mining procedures to perform our bibliometric analysis. Our application of these procedures provides the data used to perform the stages of our analysis as specified in our expert system framework for analyzing innovation through patent data shown in Fig. 2. In what follows, we will briefly describe the framework, its relation to traditional bibliometric analysis, and our implementation. For greater detail on the implementation of specific text or data mining procedures refer to Appendix B.

Using text mining tools, we extract textual content from patent application titles in order to quantify patent application activity over time (Stage 1), identify topical areas of interest and patterns

[1] http://www.forbes.com/companies/sap/.

Table 2
SAP industry strategy solutions (http://www.sap.com/industries/index.epx).

Aerospace and defense	High tech	Oil and gas
Automotive	Higher education and research	Professional services
Banking	Industrial machinery and components	Public sector
Chemicals	Insurance	Retail
Consumer products	Life Sciences	Telecommunications
Defense and security	Media	Transportation and logistics
Engineering, construction, and operations	Mill products	Utilities
Healthcare	Mining	Wholesale distribution

Table 3
Frequency of patent applications and word roots per year.

Year	Number of patents	Words count > 5
2002	1	
2003	4	
2004	20	method(s) - 12, system(s) - 14
2005	90	comput(ation, ed, er, erized, ing) - 5, defin(ed, ing, ition) - 5, promotion - 5, auto(mated, matic, matically, mation) - 6, sale(s) - 6, allocat(ing, ion) - 7, product(s) - 8, assortment - 9, us(age, e, ing) - 9, data - 11, manag(ed, ement, er, ing) - 11, pric(e, ing) - 11, process(es, ing) - 11, order(s, ing) - 13, purchas(e, ing) - 14, plan(ned, ning) - 15, method(s) - 68, system(s) - 68
2006	77	application(s) - 5, file(s) - 5, generat(e, ed, ing, ion, or) - 5, integrat(ed, ing, ion) - 5, interface(s) - 5, support(ed, ing) - 5, us(age, e, ing) - 5, web - 5, process(es, ing) - 6, service(s) - 7, data - 12, manag(ed, ement, er, ing) - 13, method(s) - 43 system(s) - 43
2007	222	access(ing) - 5, commerce - 5, configur(able, ation, ing) - 5, employ(ee, ing, ment) - 5, graphic(al, ally) - 5, map(ping) - 5, site - 5, source(s) - 5, support(ed, ing) - 5, tool(s) - 5, execut(able, ing, ion) - 6, seller's - 6, environment(s) - 7, integrat(ed, ing, ion) - 7, monitor(ing) - 7, order(s, ing) - 7, web - 7, control(led, ling, s) - 8, dynamic(ally) - 8, multi(ple) - 8, apparatus(es) - 9, auction(s) - 9, business - 9, context - 9, enterprise - 9, model(s, ing) - 9, table(s) - 9, comput(ation, ed, er, erized, ing) - 10, generat(e, ed, ing, ion, or) - 10, product(s) - 10, interface(s) - 11, information - 13, software - 14, auto(mated, matic, matically, mation) – 16, object(s) – 16, user(s) - 16, application(s) – 17, manag(ed, ement, er, ing) – 17, data – 20, service(s) – 20, base(d) – 21, provid(er, ers, ing) – 21, process(es, ing) – 24, system(s) – 118, method(s) - 120
2008	305	architectur(al, e, es) - 5, compos(ing, ite, ition) - 5, exchang(e, ing) - 5, monitor(ing) - 5, oriented - 5, present(ation, ing) - 5, server(s) - 5, tool(s) - 5, virtual(ized) - 5, access(ing) - 6, cent(er, ral, ralized, rally, ric) - 6, common - 6, configur(able, ation, ing) - 6, develop(ing, ment) - 6, document(ation, s) - 6, enhanc(ed, ing) - 6, implement(ation, ations, ed, ing) - 6, network(s, ed) - 6, pattern(s) - 6, search(able, ing) - 6, structur(e, ed, es, ing) - 6, time - 6, auto(mated, matic, matically, mation) - 7, client - 7, content(s) - 7, display(ing) - 7, evaluat(ing, ion) - 7, integrat(ed, ing, ion) - 7, support(ed, ing) - 7, web - 7, xml - 7, communication(s) - 8, context - 8, distribut(ed, ion) - 8, dynamic(ally) - 8, generic - 8, messag(e, es, ing) - 8, plan(ned, ning) - 8, test(ing) - 8, trac(e, es, ing) - 8, comput(ation, ed, er, erized, ing) - 9, product(s) - 9, database(s) - 10, framework - 10, generat(e, ed, ing, ion, or) - 10, multi(ple) - 10, control(led, ling, s) - 11, enterprise - 11, us(age, e, ing) - 13, model(s, ing) - 14, provid(er, ers, ing) - 14, service(s) - 14, information - 15, user(s) - 15, apparatus(es) - 18, software - 18, base(d) - 19, business - 19, interface(s) - 19, process(es, ing) - 22, application(s) - 23, object(s) - 27, manag(ed, ement, er, ing) – 31, data – 46, system(s) – 112, method(s) - 122
2009	194	access(ing) - 5, auto(mated, matic, matically, mation) - 5, consisten(cy, t) - 5, environment(s) - 5, human - 5, program(s) - 5, supply - 5, support(ed, ing) - 5, valid(ate, ating, ation, ity) - 5, ware - 5, architectur(al, e, es) - 6, base(d) - 6, context - 6, enhanc(ed, ing) - 6, execut(able, ing, ion) - 6, implement(ation, ations, ed, ing) - 6, multi(ple) - 6, secur(e, ely, ing, ity) - 6, database(s) - 7, dynamic(ally) - 7, framework - 7, information - 7, interface(s) - 7, item - 7, product(s) - 7, heterogeneous - 8, item(s) - 8, structur(e, ed, es, ing) - 8, table(s) - 8, comput(ation, ed, er, erized, ing) - 9, provid(er, ers, ing) - 9, apparatus(es) - 11, software - 12, us(age, e, ing) - 12, application(s) - 15, service(s) - 15, business - 16, object(s) - 17, data - 18, process(es, ing) - 20, model(s, ing) - 22, manag(ed, ement, er, ing) - 27, method(s) – 77, system(s) - 77
2010	251	cent(er, ral, ralized, rally, ric) - 5, custom(izable, ization, ized, izing) - 5, defin(ed, ing, ition) - 5, determin(ation, ing) - 5, device(s) - 5, distribut(ed, ion) - 5, event(s) - 5, handling - 5, hierarch(y, ies, ical) - 5, messag(e, es, ing) - 5, monitor(ing) - 5, perform(ance, ing) - 5, quer(ies, y) - 5, reusable - 5, structur(e, ed, es, ing) - 5, type(s) - 5, analy(ses, sis, tical, tics, zing) - 6, code - 6, document(ation, s) - 6, dynamic(ally) - 6, enhanc(ed, ing) - 6, implement(ation, ations, ed, ing) - 6, program(s) - 6, support(ed, ing) - 6, time - 6, communication(s) - 7, network(s, ed) - 7, server(s) - 7, transaction(s, al) - 7, auto(mated, matic, matically, mation) - 8, environment(s) - 8, framework - 8, product(s) - 8, resource(s) - 8, apparatus(es) - 9, user(s) - 9, base(d) - 10, design(er) - 10, provid(er, ers, ing) - 10, architectur(al, e, es) - 11, configur(able, ation, ing) - 11, database(s) - 11, enterprise - 11, model(s, ing) - 11, information - 12, business - 13, generat(e, ed, ing, ion, or) - 13, interface(s) - 13, secur(e, ely, ing, ity) - 13, multi(ple) - 16, object(s) - 18, service(s) - 18, process(es, ing) - 19, us(age, e, ing) - 20, comput(ation, ed, er, erized, ing) - 22, manag(ed, ement, er, ing) - 23, software - 28, application(s) – 35, data – 42, method(s) – 103, system(s) - 117
2011	212	analy(ses, sis, tical, tics, zing) - 5, communication(s) - 5, component(s) - 5, configur(able, ation, ing) - 5, dynamic(ally) - 5, information - 5, parameter(s) - 5, resource(s) - 5, rule(s) - 5, structur(e, ed, es, ing) - 5, tool(s) - 5, valid(ate, ating, ation, ity) - 5, web - 5, database(s) - 6, quer(ies, y) - 6, consisten(cy, t) - 7, event(s) - 7, execut(able, ing, ion) - 7, heterogeneous - 7, provid(er, ers, ing) - 7, support(ed, ing) - 7, value(s) - 7, access(ing) - 8, enterprise - 8, framework - 8, integrat(ed, ing, ion) - 8, perform(ance, ing) - 8, secur(e, ely, ing, ity) - 8, control(led, ling, s) - 9, comput(ation, ed, er, erized, ing) - 10, multi(ple) - 10, us(age, e, ing) - 11, user(s) - 12, generat(e, ed, ing, ion, or) - 13, object(s) - 13, base(d) - 15, model(s, ing) - 15, service(s) - 17, interface(s) - 22, business - 23, process(es, ing) - 24, manag(ed, ement, er, ing) - 28, application(s) – 31, data – 33, method(s) – 78, system(s) - 92
2012	322	cloud - 5, code - 5, compliance - 5, configur(able, ation, ing) - 5, custom(izable, ization, ized, izing) - 5, document(ation, s) - 5, function(s, ality) - 5, generic - 5, hierarch(y, ies, ical) - 5, item - 5, item(s) - 5, log(ging) - 5, mobile - 5, monitor(ing) - 5, operation(s, al) - 5, oriented - 5, perform(ance, ing) - 5, plan(ned, ning) - 5, report(s, ing) - 5, resource(s) - 5, rule(s) - 5, semantic(ally) - 5, social - 5, task(s) - 5, tool(s) - 5, web - 5, control(led, ling, s) - 6, design(er) - 6, develop(ing, ment) - 6, environment(s) - 6, guid(e, ed) - 6, handling - 6, modul(e, es, ar) - 6, multi(ple) - 6, shar(e, ed, ing) - 6, structur(e, ed, es, ing) - 6, table(s) - 6, time - 6, upgrad(e, ing) - 6, access(ing) - 7, context - 7, display(ing) - 7, dynamic(ally) - 7, apparatus(es) - 8, calculat(ing, ion) - 8, engine - 8, event(s) - 8, execut(able, ing, ion) - 8, analy(ses, sis, tical, tics, zing) - 9, architectur(al, e, es) - 9, component(s) - 9, optimiz(ation, izing) - 9, search(able, ing) - 9, enterprise - 10, framework - 11, information - 11, secur(e, ely, ing, ity) - 11, test(ing) - 11, auto(mated, matic, matically, mation) - 12, distribut(ed, ion) - 12, network(s, ed) - 12, comput(ation, ed, er, erized, ing) - 13, provid(er, ers, ing) - 13, software - 13, database(s) - 15, generat(e, ed, ing, ion, or) - 16, us(age, e, ing) - 18, integrat(ed, ing, ion) - 19, interface(s) - 19, model(s, ing) - 20, user(s) - 23, application(s) - 24, service(s) - 24, object(s) - 26, manag(ed, ement, er, ing) - 27, base(d) – 35, data – 40, business – 42, process(es, ing) – 46, method(s) – 95, system(s) - 107
2013	26	consisten(cy, t) - 8, heterogeneous - 8, interface(s) - 8, manag(ed, ement, er, ing) - 9, system(s) - 9, object(s) - 10, business - 12

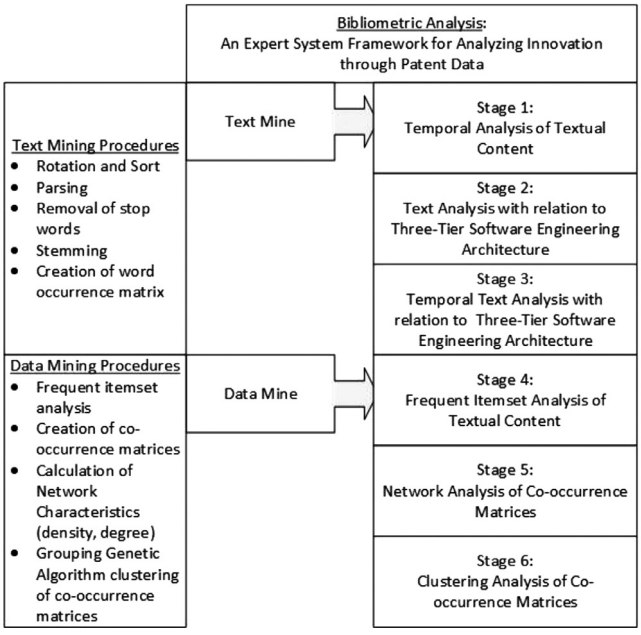

Fig. 2. Expert system framework for analyzing innovation through patent data.

(Stage 2), and illustrate temporally the trends discovered (Stage 3). These are methods of bibliometric analysis, which has been applied in many domain areas to isolate topical areas and trends. For example, Ding, Chowdhury, and Foo, 2001used it to identify patterns in the literature on information retrieval and Li, Ding, Feng, Wang, and Ho (2009) looked for trends in stem cell research; both utilized article titles in their work. Patent titles share the same richness of academic literature and are therefore an appropriate level of analysis and are consequently implemented in this study.

Bibliometrics consists of quantitative methods to analyze literature (De Bellis, 2009) and has been applied to patents, as patents are a form of publication (Gupta & Pangannaya, 2000; Daim et al., 2006). It often focuses on co-authorship networks, as well as text frequency and co-occurrence. Bibliometrics may employ text analysis strategies to examine word frequency and co-occurrence (see, for example, De Bakker, Groenewegen, and Den Hond, 2005). In Stage 4, we apply frequent itemset mining, a data mining procedure on the textual content to look for frequently occurring sets of words over all patent titles. We did not examine co-authorship since our intention is to examine innovation within a company rather than surrounding a topical area. However, co-authorship is often examined by using network analysis that demonstrates the relationships between various authors. We adopt this approach, but apply it to the text rather than the authors. In Stage 5, we calculate and examine the network metrics for a co-occurrence matrix created from the parsed text data. We then explore the relationships between the words, that is, identify word groupings that share patent application titles (Stage 6).

The framework in Fig. 2 illustrates the procedures applied to manipulate the data for each stage of the analysis. First, we applied the rotation and sort procedure (Conlon, Lukose, Hale, & Vinjamur, 2007) to all the patent application titles to identify phrases. We then used code written by the research team to parse the patent application titles, remove the stop words (see Appendix B), and manually stemmed the words. Our resulting data set contained $n = 653$ unique roots from $m = 1724$ patent application titles. The code provided the data set as an (m, n) matrix called a word occurrence matrix, where a 1 in cell (m, n) denotes that title m contains word root n. The data resulting from the application of these procedures allowed us to perform the analyses in Stages 1, 2, 3, and 4 of the framework, the results of which will be presented in Section 4. Using the word occurrence matrix, we created a co-occurrence matrix (n, n), where a 1 in cell (x, y) denotes that word x is in at least one patent application title with word y. We then ran a grouping genetic algorithm (James, Brown, & Ragsdale, 2010) on the co-occurrence matrix to obtain clusters of words that appear in common patent application titles. These data sets support the analyses in Stages 5 and 6 of the framework. Please refer to Appendix B for more detail on the application and workings of the procedures. In the next section, we will provide a discussion of the analyses and results from Stages 1 through 6 of the framework.

4. Bibliographic analysis

In Stage 1, we performed a temporal analysis of the textual content. Table 2 shows that SAP's patent application activity did not really begin to be prominent until 2005. From 2007 to 2012, the number of patent applications reported for SAP has been approximately 200 a year. The number of patent applications and the prominent keywords from the titles are shown in columns 2 and 3 respectively.

Fig. 3 gives a secondary illustration of the temporal analysis, by presenting meaningful keywords from the three tiers and the supporting technologies. The figure illustrates their introduction and the years thereafter they also appeared in significant number. In 2005, a focus on traditional business areas can be seen: sales/pricing and purchasing/ordering. In 2006, we see the introduction of a focus on user interfaces, which continues until 2012 (the last year for which significant data can be obtained); this is a top tier emphasis that can also be seen in Fig. 4 (which maps the entire data set over all years to the tiered software architecture). Other logic tier keywords that appear include ecommerce, seller, auction, mapping (2007), supply (2009), transaction, analysis (2010), semantic, and social (2012). This shows a correspondence to what we know about business themes and the history of SAP described in Section 2 and Appendix A. At the beginning, SAP's patent activity shows some focus on the core business concepts that often form the base of an enterprise system. By the mid-to-late 2000's, the business world was seeing the quick introduction of web-enabled e-commerce systems and the temporal analysis reflects that in SAP's patent activity. We see the keyword enterprise appear in 2007 and remain relatively constant. In Fig. 4, we see that this is typically part of the phrases "enterprise service", "enterprise software", "enterprise system", or "enterprise resource planning". By the late 2000's a focus on supply chain management can be seen in the industry, which appears around this time in SAP's patent activity as well. The current business environment is heavily promoting analytics and transaction level analysis, as well as social applications and text mining. In 2012, we see some of this influence reflected in SAP's patent activity. This analysis lends credence to the suggestion of business world demands influencing SAP's innovation.

The data tier is represented in this analysis as well. The prominence of the keyword database begins in 2008. We also see the introduction of the keyword cloud in 2012, which is a very popular storage option for many companies at present. Topics in all of the primary supporting technology areas are represented in Fig. 3. The auditing functions "trace" and "monitor" appear relatively early in the timeline. Access control and security, which represent the methods to protect confidentiality and integrity of data figure predominately in the timeline. Platforms and connectivity keywords are present: server, web, mobile, and network. The introduction of "mobile" in 2012 is interesting as there has been a very recent push towards migrating software functionality to mobile platforms. This trend appears to be customer driven, in that large-scale use of mobile platforms by users are making it very attractive for businesses to provide mobile solutions. Programming concepts also appear in the introduction of xml and event handling.

Over all years, there were 1724 patents. Out of these patents, 653 unique, meaningful word roots were extracted. Only 100 of the word roots were used in 17 or more patent applications titles

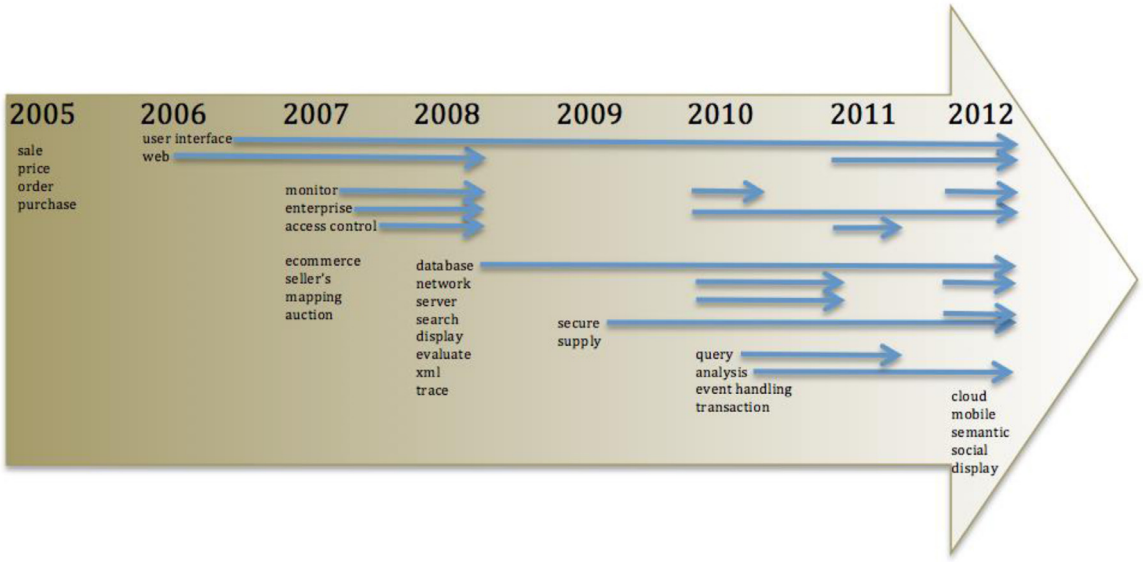

Fig. 3. Timeline of key concepts from SAPs patent applications.

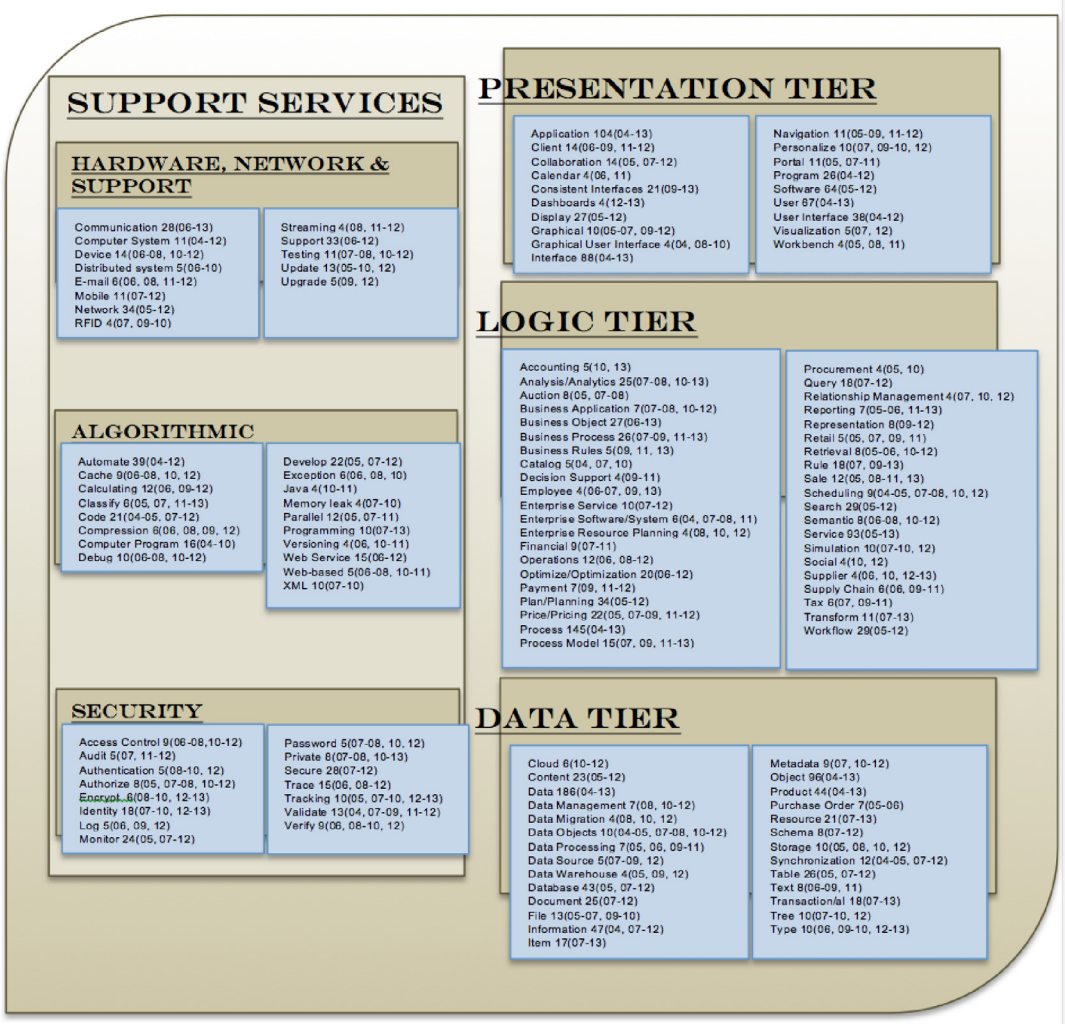

Fig. 4. Text analysis organized under three-tier software engineering architecture [count(years)].

(approximately 1% of the titles). The two-word roots that appeared most frequently were system(s) and method(s), which are not very descriptive. These two-word roots were present in approximately 40% of the titles. Interestingly, the next most frequent word, data, appeared in only 230 titles (approximately 13%). There were only 11 roots that appeared in more than 100 titles. This indicates that word root occurrence matrix is sparse. That is, the words were not frequently reused in the titles. This indicates that SAP may have quite a bit of variety in their areas of innovation, but little repetition.

In Stage 2, we determined that the textual content from SAP's patent activity could be mapped to the three-tier software engineering architecture to assist in interpretation. The three-tier software engineering architecture separates software into three conceptual levels. In software engineering the business logic layer is often referred to as the middle tier (logic tier or application tier) of the three-tiered architecture (Froese, Yu, Liston, & Fischer, 2000). In an enterprise system, this is where the logic for the primary business functions (e.g., inventory and order management) resides. The top tier (presentation tier) revolves around the user interfaces to the system. In this tier, the emphasis is on usability of the system and presentation in formats that enhance the user experience and the usefulness of the system. The bottom tier is the data tier. This tier is typically a data management system of some kind. The data tier is especially important to enterprise systems because of the large volume of data that businesses generate and especially appropriate due to a corporate focus on data analytics.

Software solutions often depend on other technology to function well. For enterprise systems, these supporting technologies include: hardware, security solutions (integrated or added-on), networking services, programming services, data management (software and hardware) platforms, and testing, implementation, and support functions. These supporting technologies are needed to run the software system (servers, storage), required to let businesses utilizing the software systems communicate among themselves through the software (networking), necessary to protect the integrity and confidentiality of data collected and maintained (security), required to store and manage the data collected (database management systems, cloud services), and necessary to support development (programming language). These are just a few examples, but for complex software systems, these supporting technologies are plentiful and quite integral. The tight coupling of the software systems with these supporting technologies is likely to drive familiarity and therefore innovative activity that integrates these technologies.

Fig. 4 illustrates the results of the text analysis of all patent application titles for all years in the data set (2001,Feb. 2013) mapped to the three-tier software engineering architecture. In this figure, meaningful words that reoccur in the patent application titles have been identified from the text analysis and categorized based upon their relation to the three-tier architecture and the supporting technologies. Next to each term, the frequency with which it appears in the titles is displayed, along with the years in which it appears in parentheses. The rotation and sort procedure provides the ability to extract phrases, which can be seen in Fig. 4 as well. These terms were manually extracted using the following criteria: the word or phrase had to occur in four or more patent application titles, be meaningful, and be relevant to one of the elements of the classification (a tier or a supporting category). By manually extracting the words and phrases at this stage, we were able to use context to categorize the term within the architecture. In other words, we were able to use the full or partial titles to provide further context for the classification. As can be seen from examining Table 2, not all frequently occurring words are shown in Fig. 4. Only those that were meaningful could be appropriately classified. Approximately 30 2-word phrases were identified. In the case of frequently occurring 2-word sets, the individual words are also part of the totals given to the side in the instances where the single word is classified. We were able to locate only one repeating, meaningful phrase containing over two words, "Enterprise Resource Planning", which is the category of software. It is interesting that in the patent titles, this phrase occurs only four times. The frequent itemset analysis by definition will find all frequent 2-item subsets, including phrases (although it will not identify them as phrases). While we will see below that we used a higher threshold for occurrence in the frequent itemset analysis, we could use the output to double-check that the manual process did not miss any frequently occurring k-word phrases.

It can be seen in Fig. 4 that the quantity of meaningful terms is quite plentiful. However, the counts (frequencies) also illustrate a low amount of repetition in use. This indicates that SAP's innovation has breadth, but not much depth. The grouping of the terms into the framework presented in Fig. 4, allows us to gain a visual appreciation of the spread of the activities over the tiers and the supporting technologies. It is shown that SAP is actively exploring interface design activities, which is evidenced by the repetitious use of "user interface", "graphical user interface", and "consistent interface". The idea of presenting a consistent interface to users is a specific design requirement. Visualization, personalize, etc. describe active efforts to present data in a useable way. The business logic tier illustrates activities in most major business functions: retail, sales, procurement, financial, accounting/tax, payment, scheduling, supply chain, customer relationship, etc. Also present is the basic logic underlying most computerized processes: query, optimization, simulation, retrieval, reporting, search, decision support, etc. The data tier is also well covered, with major types of data and representations: data object, file, table, schema, metadata, tree, etc. Storage is represented as well: database and data warehouse. The same trend of broad coverage is apparent in the supporting technologies. Hardware systems and platforms are recognizable: computer system, distributed system, device, and mobile. One interesting technology that appears in this category is RFID (radio frequency identification). This hardware is being widely implemented for inventory management/supply chain tracking. Networks, e-mail, and communication designate a presence of networking technologies. Streaming also appears which is necessary for video applications. Update, upgrade, and testing are support areas that would be expected for a software development company. Notable mentions of particular programming languages, Java and XML, appear. Typical programming concerns are present in versioning, cache, calculating, classify, and debugging. An emphasis on web-based and web service is important, as many ERP companies are rapidly moving toward SaaS and web-based platforms. The combination of distributed system and parallel show some migration to parallel computing, which is often necessary for large-scale data mining. All major areas of security are present. Password, access control, authentication, authorize and password all fall within identity and access management. Encrypt is the major function for confidentiality (private). Also not surprising for a data intensive area is an emphasis on audit keywords: verify, validate, trace, tracking, monitor, and log. Audit technologies help provide data integrity, among other uses. Again, the analysis points to wide-ranging innovation at SAP and quite a bit of integration of the supporting technologies, along with broad activity in the tiered architecture model.

In Stage 3, we examined the mapping of the textual content to the three-tier software engineering architecture over time. Figs. 5 and 6 summarize the activity over the tiers and the supporting technologies over time, providing a temporal element to the classification. Fig. 5 explores the occurrence in patent applications of terms classified in each tier of the architecture each year. The figure illustrates the number of patents each word in that tier appeared in that year. It can be seen that in most years, terms from the data tier appeared with more frequency than from the other two tiers, followed by the logic tier and the presentation tier. This is interesting in that it shows a focus on the data management processes rather than the business logic. That the backend of the enterprise system is one of the most important

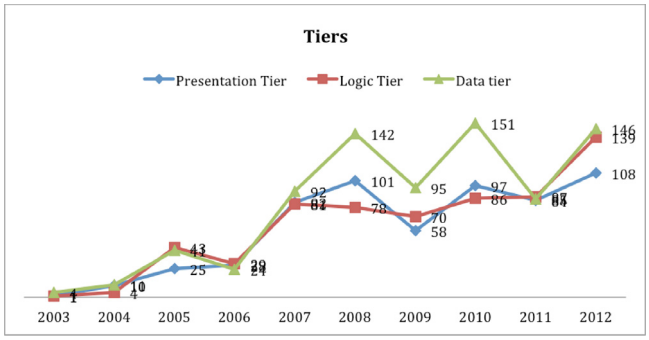

Fig. 5. Occurrences of words from each tier in patent applications by year.

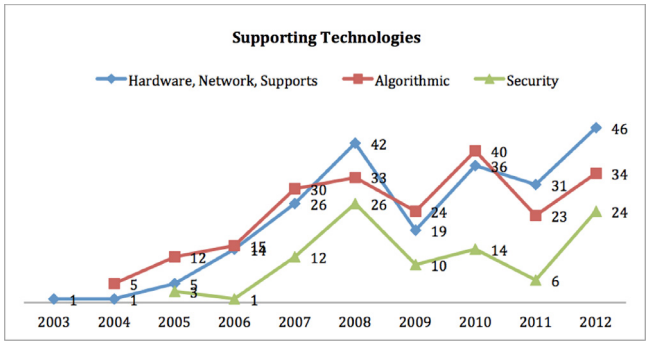

Fig. 6. Occurrences of words from each supporting technology category in patent applications by year.

components cannot be argued, but the emphasis on terms in this category where it could be suggested there is less breadth than in the logic tier is interesting. Fig. 6 provides the same view of the supporting technologies. Terms from the algorithmic category are the most frequent for a majority of years, followed closely by terms in the hardware category. That these two categories have the most focus is not surprising, as they have been historically very closely tied to software, they provide the development environment and logic and the platforms to run it. Security has often been an afterthought in software development, although in recent years, there has been a push to integrate security more tightly into the software. That it is well represented in SAP's patents is encouraging.

In Stage 4, we explored the output of the frequent itemset procedure. The frequent itemset analysis using the Apriori algorithm revealed no meaningful itemsets other than those already identified in the earlier text analysis. For the Apriori algorithm to obtain a significant number of frequent itemsets, the support threshold had to be reduced to 1.1%. Support is defined as the minimum percentage of all patent application titles the itemset must appear in to be deemed frequent. At 1.1%, a pattern must appear in approximately 19 titles to be deemed frequent. Less than 1.1% became cumbersome for the algorithm. We manually reduced the support in the previous analysis. The frequent itemset analysis looks for frequent patterns of co-occurrence only. That is, subsets of words that appear frequently in the same titles, regardless of proximity to each other in the text. It is interesting that other than some of the phrases already identified in the previous analysis, relaxing the requirement of appearing consecutively did not generate any meaningful new groups of words appearing in the same titles. However, this result does reinforce the conclusion of breadth of innovation rather than intensity in a few focal areas.

In Stage 5, we examined the results of the network analysis, which are illustrated in Figs. 7 and 8. Fig. 7 visualizes the relationship network between the word roots. The nodes represent the 653 word roots and a line connecting one node with another designates that those two nodes appear in at least 1 patent application title together. In an undirected graph representing a square matrix of (653, 653), there is the possibility of 212,878 arcs. The network in Fig. 7 has 20,864 arcs (relationships). Density is a quantitative measure of how densely connected the graph is that is calculated by dividing the number of arcs present in the graph by the number possible. The density of the graph in Fig. 7 is 0.098, which quantitatively shows that the graph is not very dense. The interpretation is that SAP's patent application titles are not strongly related, which reaffirms once again the finding that SAP innovates in many areas, but that they do not show frequent repetitious topical areas. Degree denotes the number of arcs emanating from a node. The average degree provides an idea of how well connected a node is, which translates to how many other terms it appears in patents with. The average degree of the network appearing in Fig. 7 is 31.9, which shows that relaxing the co-occurrence frequency requirement indicates some relatedness. Over all years, a word root shares at least one patent application title with on average 31 other word roots.

In Fig. 7, nodes of the same shape appearing in a bunch denote a cluster. The clusters are a result of the application of the grouping genetic algorithm, which provides the data for Stage 6. While not encouraged, the algorithm allows for small groups, even groups of only one word. Fig. 8 displays some of the larger groups mapped to the three-tier software engineering architecture. Groups that were not meaningful enough to permit classification or that were very small were excluded from the figure. The groups from Fig. 8 are also shown in Fig. 7. As previously mentioned, this clustering reduces the restriction of frequency in co-occurrence. So these clusters allow for a relaxed relationship between the members (word roots). They are related through co-occurrence, but more loosely than in the frequent itemset analysis. However, this analysis allows for more descriptive groupings of words. For example, in both figures, the group [supply, chain, encrypt(ed, ion), item(s), rfid, tag(ging, s), trac(e, es, ing)] is shown. RFID is often used as a way to trace inventory (items) through a supply chain, that encryption is included shows a security supporting technology, combined with the hardware supporting technology (RFID) and a business logic area (supply chain). The clusters provide a more expansive context of the activity. While we have determined in previous analysis that repetition in SAP's innovation topics is not high, but the scope is quite wide, the cluster analysis allows us to examine the scope of the innovation with a little more context. The number of clusters and the spread over the framework again illustrates the large scope of their innovation.

The clusters provide more depth to the areas already identified. We already have identified a focus on authorization in the security area. The clustering algorithm extracts the group (third, party, certification), which is a specific method. Thus, Fig. 8 provides a loose overview of SAP's innovation areas. One notable result of this analysis is to examine the data tier of Fig. 8. The clusters that appear in the data tier illustrate many of the current topics in data mining: (aggregating, key, characteristics), (evaluating, constraint, patterns), (similarity, search, retrieval, capability), (correlation, decision, real, synchronization), (aided, developing, extraction, machine, rule), and (behavior, cache, compression, improved, scheme). Popular topics in data mining include: compression, pattern finding using constraints, correlation analysis, evaluating patterns found in data, automatic rule extraction, etc. The number of large groups that can be mapped to data mining shows an interesting emphasis on data mining techniques, which reinforces the earlier finding regarding the importance of the data tier.

5. Discussion and concluding remarks

Fig. 9 presents a high-level summary of our findings from our bibliometric analysis of SAP's patent activity using our expert system framework for analyzing innovation through patent data. In Stage 1 we saw that SAP's patent activity started in earnest in 2005. The

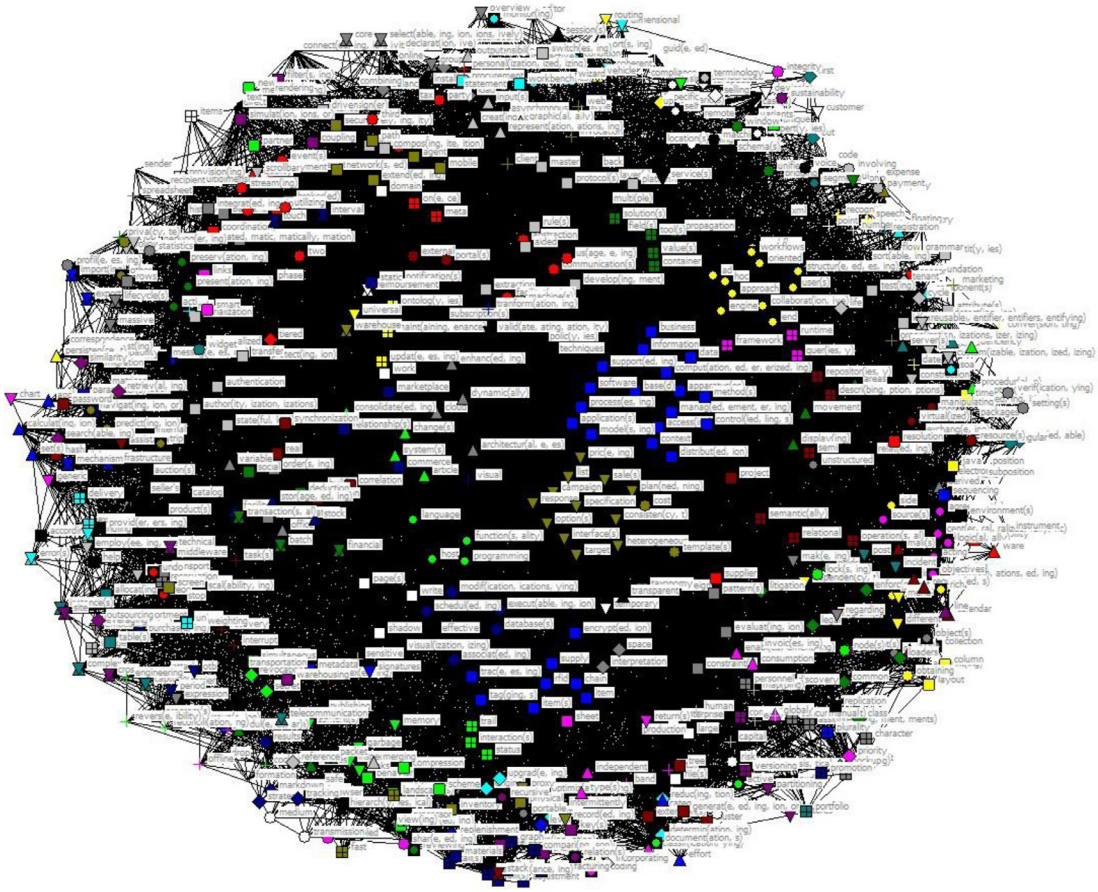

Fig. 7. SAP patent word clusters (654 words from 1724 patent application titles).

complexity of enterprise systems with their heavy dependence on supporting technologies would suggest that the scope of technological areas over which an enterprise company could be innovative would be large. The results of Stages 2 and 3 indicate that SAP is very actively innovating in a broad array of areas. The data management tier receives a great deal of focus in SAP's innovative activity. It is also illustrated that emerging technologies are quickly integrated into SAP's innovation. The analyses in Stages 4 and 5 further illustrate that SAP's innovation has significant breadth (i.e., work is being performed in many areas) but little depth (i.e., there is not much topical repetition). Findings from Stage 6 illustrate an emphasis on data mining focus areas, as well as security and other emerging technology focus areas.

Previous research has suggested that patent analysis can provide many beneficial outcomes, including: detecting trends in innovation, forecasting emerging technologies, and assisting companies in determining their own strategic opportunities or assessing their competitors' strategies. Patent analysis could also be used as part of determining the health or status of a company to assist in merger and acquisition decisions or large-scale capital purchases where the health and innovation status of the company is important in assessing access to future product development and support.

SAP provided an interesting case study for our framework because it is a market leader in its software category, produces a software system extremely integral to the functioning of corporations, and solely focuses on this one software category. Our analysis of SAP illustrates that the company is actively innovating in across the three-tier software engineering architecture and supporting services. We show that SAP's patent activity has grown and shows a focus on emerging technologies in data mining and security, among others. The analysis shows that initial focus on core business logic was later followed by quick integration of new technologies, which could indicate forethought and a strategy to stay current with technological trends in ways that can best be integrated into their core business product. SAP's patent activity shows significant focus on data management for many years. Considering that analytics has become an increasingly visible business concern, SAP's activity in this area is timely and shows an awareness of the importance of utilization of the data their systems collect.

Our findings serve to present a corporate-level analysis of innovation trends that could be used by decision-makers within the company itself to consider its innovation or R&D strategies. Alternatively, smaller enterprise software companies could use this analysis to consider their own innovation strategy as they attempt to move further into this software space. For example, a company could look at SAP's broad innovation and quick exploration of emerging technologies and follow a similar strategy. Alternatively, it could use this type of analysis to determine an under-represented niche that a company could fill or an opportunity for a company to develop software to supplement a system like SAP provides. Last, companies considering the large-capital purchase that a SAP installation represents could consider this analysis to assist in decision-making. For example, a company considering purchasing SAP could be looking for a software implementation that keeps up with emerging technology trends or a company with great depth in innovation. However, these characteristics may mean a lot of software changes and re-training, as well as a depth that is not necessary or desired for others.

However, while our analysis can provide insight into the strategy of the recognized leader of enterprise resource planning software (enterprise systems software), one limitation to our choice of company to analyze is its unique position in the competitive landscape. We identified 381 enterprise software companies, other than

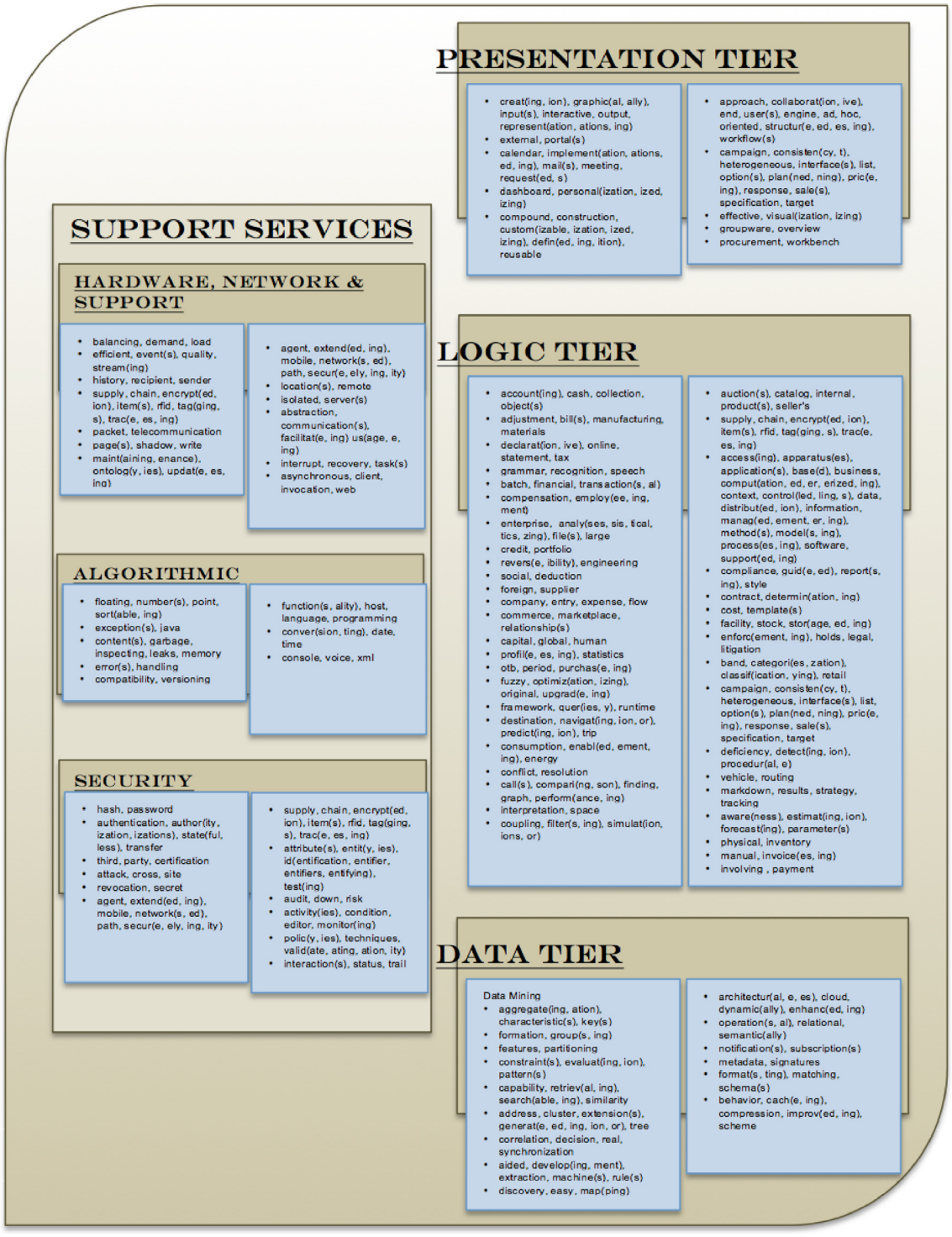

Fig. 8. Network analysis organized under three-tier software engineering architecture.

SAP, that focused solely on enterprise systems. SAP held 1724 patents and all other 381 companies held less than 1000 patents combined. This means that, as of our data collection, SAP dominates the patented innovation space of sole-enterprise system software developers. However, several very diversified companies also provide commercial enterprise software, including: Microsoft, Oracle, Rockwell, GE Industrial Solutions, IBM, Fujitsu, Hewlett-Packard, Honeywell, and Siemens. The patents submitted by these companies for their enterprise management software would not be distinguishable from the patents for their other products and software systems. Innovation strategies for this variety of diversified personal and corporate software providers, hardware providers, and industrial systems companies would likely be very different from a company focused solely on enterprise systems software. This competitive landscape would not necessarily be seen in other industries to which this framework could be applied, in which case, the analysis could be done on multiple companies to compare corporate strategies (i.e., one of the outcomes suggested above). Future work could include applying our framework to analyze and contrast multiple companies' innovation strategies.

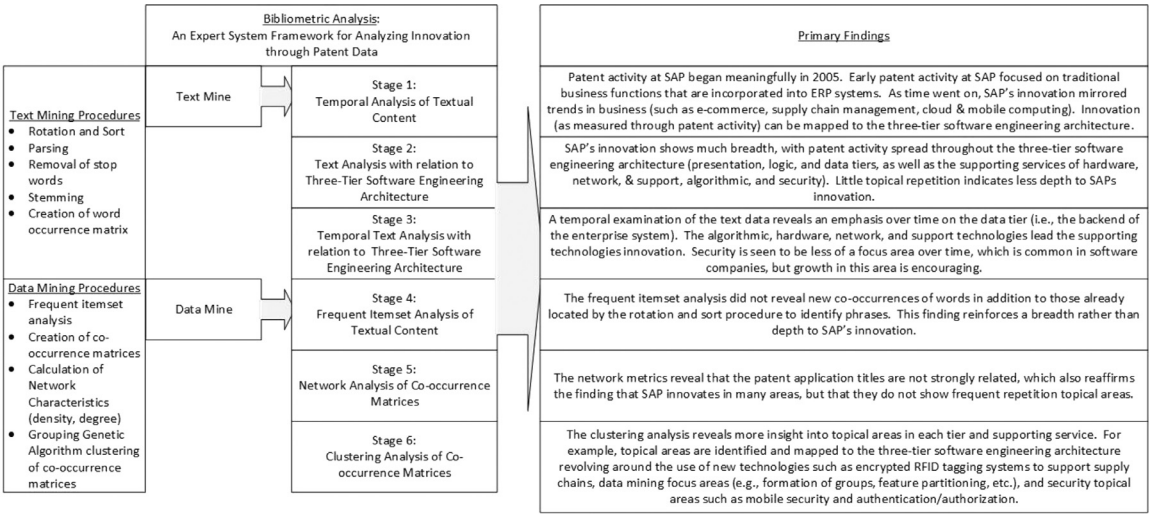

Fig. 9. High-level summary of findings from application of framework to SAP patent data.

Our expert system framework for analyzing innovation through patent data provides a flexible integration of text and data mining tools to provide temporal and trend analyses of textual content. We borrow from text mining and network analysis techniques to process data for six stages of analysis. Our framework differs from other work on patents in the expert systems area by utilizing a different combination, and in many cases a more robust set, of techniques. For example, Shih et al. (2010) combined a step to calculate traditional patent indicator variables with a decision-rule data mining technique to explore trends over the semiconductor industry. Where our approach focuses on textual content, this approach focuses more on numeric data. Thus, our framework relies on text mining and network analysis techniques that are not applied in the work of Shih et al. (2010). Future work could include integrating such numeric variables into our framework or creation of numeric variables from our textual content. As another example, Kim et al. (2008) utilize textual content, but with a different approach. Rather than starting by extracting textual content, their study employs experts to determine keywords. They then use those keywords to search for patents and those patents are used to structure semantic networks of keywords. Similar to part of our framework, they do use a clustering algorithm to group the patents. However, they use a k-means clustering algorithm rather than our grouping genetic algorithm, they cluster patents whereas we cluster words, and our analysis includes several stages of analyzing a word set constrained by patents from a particular company whereas their analysis focuses on creating a patent map from the clustering of patents obtained by keywords provided by experts. Thus, our approach differs but is complementary to other expert systems approaches and our framework is flexible enough to include such approaches dependent on the desired outcome of the overall analysis.

The previous examples illustrate the contributions of our approach. Our first contribution is that we present a framework for the analysis of textual content utilizing both text and data mining techniques in an organized (staged) approach. Second, our framework is flexible in that techniques could be substituted or stages added. For example, a k-means or principle components analysis could be used to create groupings rather than our grouping genetic algorithm implementation. Furthermore, stages could also be added to our framework depending on the needs of the analysis. For example, traditional bibliometric analysis sometimes includes an examination of co-authorship. This could be used, for example, to explore the impact of corporate or academic alliances over patents for an entire industry. One could include this analysis as a new stage that uses the network analysis techniques on a different complementary dataset. Third, our framework could be used to analyze textual content at different levels of analysis. That is, trends between units of a single corporation, from a single corporation (our example), between multiple corporations, or from an entire industry. Fourth, it could be used for multiple outcomes. For example, to determine trends, to study strategy, etc. Our application of the framework on SAP illustrated the examination of trends within a single corporation using a six-stage approach on only textual data (patent application titles) and applying some novel text and data mining techniques from the literature.

Our approach is not without limitations. First, our application dataset was small enough to do some interpretation by hand and to use a sophisticated clustering technique. For companies with larger patent portfolios or if the full text of the patents were used, other tools may need to be substituted into the framework to automate the mapping of the terms to the chosen topology (e.g., the three-tiered software engineering architecture) or this mapping may become more time consuming if the number of reused and meaningful terms becomes larger. Similarly, the grouping genetic algorithm clustering technique will not cluster very large datasets (>1000) without needing to be parallelized, so one may need to substitute a k-means type clustering analysis or explore partitioning or aggregation techniques to add to the framework toolset. Second, as in our application, the context may limit the outcomes that can be obtained. Our application was restricted to a single company analysis because the differences in competitors limited the usefulness of a between corporation analysis.

Future research could include using the framework to explore applications that benefit from different levels of analysis (industry, country, etc.). It may also be interesting to combine numeric analyses into the framework and/or create numeric variables from some stages of the framework to complement a different type of analysis (such as rule-based association mining) or a predictive model in a new stage. In addition, more text could be consumed into the processes (e.g., abstracts, citation titles, full patent text) for different outcomes. Related to these suggestions, various outcomes could be explored by applying the framework to different subsets of patents. For example, patents for an industry (e.g., semiconductors and pharmaceuticals) could be explored in this fashion to determine industry trends or similarly innovating companies within an industry.

Appendix A

SAP describes the history of their company in four blocks of time: 1972–1981: the early years; 1982–1991: the SAP R/2 Era;

1992–2001: the SAP R/3 Era; and 2002–Present: Real-Time Data Where and When You Need It (http://www.sap.com/corporate-en/our-company/history/index.epx). The first major version of SAP was known as SAP R/1 and was financial and accounting software that included three components, purchasing, inventory management, and accounting, based on a single database (Leimbach, 2008). SAP R/2 was introduced in 1980 and used a two-tier architecture designed to work on mainframe computers. R/2 was expanded over time to include additional currencies and fiscal regulations and thus began to provide the infrastructure required for international trade (Campbell-Kelly, 2003). SAP opened its first US office in Philadelphia because of its proximity to multinational companies that were already using SAP in Germany, primarily in the oil, pharmaceutical, and chemical industries (Campbell-Kelly, 2003). SAP R/3, designed for client server computing, was released in 1992, with significant upgrades released in 1999, 2001, and 2003 (Inc. Kogent Learning Systems, 2010). Campbell-Kelly (2003) reports that SAP "went into hyper-growth" in 1993, at least partially due to the popularity of business process reengineering in the US. SAP's web portal offering, mySAP, was introduced in 1999 to combine e-commerce solutions with SAP's existing ERP applications to create a preconfigured system that could run on a single database (D'Amico, 1999; Inc. Kogent Learning Systems, 2010). The mySAP Business Suite includes various applications such as mySAP ERP, mySAP Supply Chain Management, and mySAP Customer Relationship Management. SAP Business Suite 7 was introduced in 2009 as the "company's next generation suite enable by service oriented architecture (SOA)" (http://www.sap.com/solutions/business-suite/newsevents/press.epx?pressid=11266). The SAP Business Suite includes "ready to run" business processes for departments within an organization (Missbach et al., 2013). The evolution of SAP software continues now with HANA, as it was introduced to selected customers in 2010 and became generally available in 2011 (SAP announces general availability of SAP HANA, 2011). HANA is SAPs in-memory high performance analytic appliance (Missbach et al., 2013) and SAP chairman Hasso Plattner describes it as the realization of a dream to "reinvent enterprise systems" (Henschen, 2013). Users of SAP will be able to simplify their information technology (IT) systems as the separation between transactional and business intelligence infrastructure has been removed in HANA (Henschen, 2013).

Appendix B

Rotation and sort procedure

We processed the textual data by applying a rotation and sort procedure (Conlon et al., 2007) that provides output of the type shown in Fig. 10. This method provides us with the capability to analyze three characteristics of the text of the patent application titles. First, we can see groupings of words. In Fig. 10, we see that the primary word of this rotation is "tax". In our analysis, we use word stems, which means that "tax" and "taxes" will be logically grouped together. Fig. 10 illustrates that the root "tax" appears at some point in 6 titles. Second, this analysis shows us the years of occurrence: the word "tax" appeared in 2007, 2009, 2010, and 2011. Third, we can identify common multi-word phrases through this analysis. For example, looking at the output in Fig. 10, we see "tax declaration" appear twice.

Parsing, stop words, stemming, and creation of word occurrence matrix

We took the patent applications and ran them through a text parser to extract the words, then combined terms with the same stem, and finally created a word occurrence matrix. This word occurrence matrix is a title (m) by word (n) matrix, in which a cell (m,n) denotes that title m, contains word n. From this output, we can also obtain word root frequencies, in the form of a count of the number of titles in which a word appears. It should be noted that to reduce the number of words, only meaningful terms were extracted from the patent titles. The parser was configured to ignore punctuation as well as stop words such as common articles, prepositions, conjunctions, and pronouns. The list of words the parser was configured to exclude from the matrix is: a, across, after, against, among, an, and, and, any, are, as, at, be, between, by, during, for, from, given, having, high, in, including, inside, inter, into, is, like, more, near, non, of, on, only, or, over, pre, same, such, sub, that, the, therefore, thereof, through, to, upon, via, when, whether, which, while, with, within, and without. Also, only words that appeared in more than one patent application title were retained for the final root word set. These refinements allow for a cleaner analysis because multiple occurrences of stop words do not mask meaningful words and singleton instances of words do not stress the clustering process.

Frequent itemset analysis

Using the word occurrence matrix, we can perform a frequent itemset analysis (Han, Kamber, & Pei 2011). A frequent itemset, in our context, is a set of words that appear together in a particular percentage of the patent application titles. Contrasted with the earlier analysis, this allows us to determine if there are frequent word groupings that appear in patents without regard to the structure of the words in the title. In other words, the first analysis lets us determine phrases (words that appear together in the structure of the title) and the frequent itemset analysis provides us with an overview of co-occurrence (groupings of words that all appear in the same set of patent titles) without regard to the semantic structure. We use the Apriori algorithm to perform the frequent itemset analysis.

Co-occurrence matrix, calculation of network characteristics, and grouping genetic algorithm clustering

Finally, we adopt network analysis tools to explore the relationships between words with regard to their occurrence in patents. Unlike the frequent itemset or text analysis, this analysis allows us to break the strict requirement that all words must appear together in the same patent application title. It gives us the ability to explore relatedness by a defined "similarity" between words. We define this relation as "word x *occurs* in at least one title with word y". With this similarity, every title that contains word x, does not have to also contain word y. They just have to co-occur at least once. We use matrix manipulation software to transform the n by m word occurrence matrix into an m by m matrix of minimal co-occurrence relations. That is, $m(x, y) = 1$ denotes that word x and word y occur in at least one patent title together. This matrix can be visualized graphically by letting the nodes of a graph represent a word root, with an arc connecting any two nodes where the relationship holds true, $m(x, y) = 1$. Visualizing

2010	Tax	Declaration	Application	Software.		
2011	tax	declaration	program.			
2007	tax	deduction.				
2009	Tax	Information.				
2011	TAX	LEGISLATION.				
2010	Taxes	In	Computer-Based	Sales	Transactions.	

Fig. 10. Output of text rotation and sort procedure.

the matrix in this way is a common network analysis tool. Using this matrix, we can also explore the relatedness of words using common network characteristics, such as degree and density. Degree of a node (a word root) quantifies how many arcs are attached to that node. In our context, this illustrates how many other words that word *x* co-occurs with in at least one title. While degree is a node level measure, density provides a network level view as a measure of the arcs in the network (the number of pairs of words that co-occur in at least one patent) to the total number of arcs that could be present in the network were it fully connected (if every word co-occurred with every other word in at least one patent). The final network analysis tool we employ is a clustering algorithm. We use a clustering algorithm on the *m* by *m* matrix to form groups of related patents. The blockmodeling algorithm is employed, which is a heuristic method written to cluster square, symmetric matrices (James et al., 2010). The goal of the algorithm is to find a small number of large groups of densely connected nodes in a graph. The clusters allow us to explore relatedness between terms to find weakly determined focal areas.

References

Adegoke, O., Walumbwa, F. O., & Myers, A. (2012). Innovation strategy, human resource policy, and firms' revenue growth: the roles of environmental uncertainty and innovation performance. *Decision Science, 43*(2), 273–302.

Ashford, W. (2011). Analysis: What is the secret sauce in SAP's success? *Computer Weekly*, [WWW document] http://www.computerweekly.com/news/2240106255/Analysis-What-is-the-secret-sauce-in-SAPs-success (Accessed 27.10.11).

Ashurst, C., Freer, A., Ekdahl, J., & Gibbons, C. (2012). Exploring IT-enabled innovation: a new paradigm? *International Journal of Information Management, 32*(4), 326–336.

Breitzman, A. F., & Mogee, M. E. (2002). The many applications of patent analysis. *Journal of Information Science, 28*(3), 187–205.

Breitzman, A., & Thomas, P. (2002). Using patent citation analysis to target/value MA candidates. *Research-Technology Management, 45*(5), 28–36.

Bos-Brouwers, H. E. J. (2010). Corporate sustainability and innovation in SMEs: evidence of themes and activities in practice. *Business Strategy and the Environment, 19*(7), 417–435.

Campbell-Kelly, M. (2003). *From airline reservations to sonic the hedgehog: a history of the software industry*. MIT Press.

Chen, R. (2009). Design patent map visualization display. *Expert Systems with Applications, 36*(10), 12362–12374.

Chen, Y. S., & Chang, K. S. (2013). The nonlinear effect of green innovation on the corporate competitive advantage. *Quality & Quantity, 47*(1), 271–286.

Collignon, S., James, T. L., & Cook, D. (2010). The future of enterprise management applications: an examination of the history and trends in practice and research. *Decision Sciences Institute Proceedings, 1*, 547–552.

Conlon, S., Lukose, S., Hale, J., & Vinjamur, A. (2007). Automatically extracting and tagging business information for e-business systems using syntactic and semantic analysis. In A. F. Salam, & J. Steven (Eds.), *Semantic web technologies and ebusiness: virtual organization and business process automation* (pp. 101–126). Idea Group Inc.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: use of bibliometrics and patent analysis. *Technological Forecasting and Social Change, 73*(8), 981–1012.

D'Amico, M. L. (1999). MySAP portal makes premiere. *InfoWorld, 21*(19), 26.

De Bakker, F. G., Groenewegen, P., & Den Hond, F. (2005). A bibliometric analysis of 30 years of research and theory on corporate social responsibility and corporate social performance. *Business & Society, 44*(3), 283–317.

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management, 37*(6), 817–842.

Forbes (2015). SAP posts strong 2014 results, cloud business to drive future growth, *Forbes*, Online [WWW document] http://www.forbes.com/sites/greatspeculations/2015/01/22/sap-posts-strong-2014-results-cloud-business-to-drive-future-growth/ (Accessed 02.08.15).

Froese, T., Yu, K., Liston, K., & Fischer, M. (2000). System architecture for AEC interoperability. In G. Gudnason (Ed.), *Proceedings of CIB-W78: international conference on construction information technology* (pp. 362–373). Icelandic Building Research Institute. June 28-30 *1*.

Gupta, V. K., & Pangannaya, N. B. (2000). Carbon nanotubes: bibliometric analysis of patents. *World Patent Information, 22*(3), 185–189.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Morgan Kaufmann.

Han, J. K., Kim, N., & Srivastava, R. K. (1998). Market orientation and organizational performance: Is innovation a missing link? *Journal of Marketing, 62*(4), 30–45.

Henschen, D. (2013). SAP moves core applications to HANA in-memory platform, *Informationweek – Online* [WWW document] http://ezproxy.lib.vt.edu:8080/login?url=http://search.proquest.com/docview/1268639066?accountid=14826 (Accessed 27.02.13).

Hicks, D., Breitzman, T., Olivastro, D., & Hamilton, K. (2001). The changing composition of innovative activity in the US—a portrait based on patent analysis. *Research Policy, 30*(4), 681–703.

Holsapple, C. W., & Sena, M. P. (2005). ERP plans and decision-support benefits. *Decision Support Systems, 38*(4), 575–590.

Inc. Kogent Learning Systems (2010). *SAP MM handbook*. Sudbury, MA: Jones & Bartlett Learning.

James, T., Brown, E., & Ragsdale, C. T. (2010). Grouping genetic algorithm for the blockmodel problem. *IEEE Transactions on Evolutionary Computation, 14*(1), 103–111.

Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications, 34*(3), 1804–1812.

Lee, S., Yoon, B., Lee, C., & Park, J. (2009). Business planning based on technological capabilities: patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change, 76*(6), 769–786.

Lee, S., Yoon, B., & Park, J. (2009). An approach to discovering new technology opportunities: keyword-based patent map approach. *Technovation, 29*(6–7), 481–497.

Leidner, D. E., Lo, J., & Preston, D. (2011). An empirical investigation of the relationship of IS strategy with firm performance. *The Journal of Strategic Information Systems, 20*(4), 419–437.

Leimbach, T. (2008). The SAP story: evolution of SAP within the German software industry. *IEEE Annals of the History of Computing, 30*(4), 60–76.

Li, L. L., Ding, G., Feng, N., Wang, M. H., & Ho, Y. S. (2009). Global stem cell research trend: bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics, 80*(1), 39–58.

Missbach, M., Stelzel, J., Gardiner, C., Anderson, G., & Tempes, M. (2013). *SAP on the cloud*. Berlin Heidelberg: Springer-Verlag.

Norton, S. (2014) Now tied to the cloud, SAP faces bumpy ride; firm is introducing online and cloud-based versions of its services. *Wall Street Journal – Online* [WWW document] June 3, 2014 http://www.wsj.com/articles/sap-faces-bumpy-ride-tied-to-cloud-web-1401838481 (Accessed 21.07.15).

Noruzy, A., Dalfard, V. M., Azhdari, B., Nazari-Shirkouhi, S., & Rezazadeh, A. (2013). Relations between transformational leadership, organizational learning, knowledge management, organizational innovation, and organizational performance: an empirical investigation of manufacturing firms. *International Journal of Advanced Manufacturing Technology, 64*(5–8), 1073–1085.

Palaniswamy, R., & Frank, T. (2000). Enhancing manufacturing performance with ERP systems. *Information Systems Management, 17*(3), 43–55.

Pullen, A. J. J., de Weerd-Nederhof, P. C., Groen, A. J., & Fisscher, O. A. M. (2012). Open innovation in practice: goal complementarity and closed NPD networks to explain differences in innovation performance for SMEs in the medical devices sector. *Journal of Product Innovation Management, 29*(6), 917–934.

SAP announces general availability of SAP HANA (2011). *Telecomworldwire* [WWW document] http://www.nfvzone.com/news/2011/06/21/5587035.htm (Accessed 27.02.13).

Schmoch, U., & Schnöring, T. (1994). Technological strategies of telecommunications equipment manufacturers: a patent analysis. *Telecommunications Policy, 18*(5), 397–413.

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.

Siwczyk, Y., Warschat, J., & Spath, D. (2012). Software-based patent analysis: how to leverage a text-mining tool. In *Proceedings of Portland international center for management of engineering and technology, PICMET: technology management for emerging technologies* (pp. 1006–1013).

Slater, I. J. (2012). To market, to market: innovation, Canada's nuclear industry, and the case of the nuclear battery. *Journal of Canadian Studies, 46*(1), 75–111.

Song, L. Z., Song, M., & Di Benedetto, A. (2009). A staged service innovation model. *Decision Sciences, 40*(3), 571–599.

Srivardhana, T., & Pawlowski, S. D. (2007). ERP systems as an enabler of sustained business process innovation: a knowledge-based view. *Journal of Strategic Information Systems, 16*(1), 51–69.

Sundbo, J. (1997). Management of innovation in services. *Service Industries Journal, 17*(3), 432–455.

Tajeddini, K., & Trueman, M. (2012). Managing Swiss hospitality: how cultural antecedents of innovation and customer-oriented value systems can influence performance in the hotel industry. *International Journal of Hospitality Management, 31*(4), 1119–1129.

Wolde, H.T. (2012). SAP software sales top forecast as wins market share. *Reuters* October 24, 2012 [WWW document] http://www.reuters.com/article/2012/10/24/us-sap-results-idUSBRE89N0CA20121024 (Accessed 25.02.13).

Yao, Y., & He, H. C. (2000). Data warehousing and the internet's impact on ERP. *IT Pro, 2*(2), 37–41.

Yoon, J., & Kim, K. (2011). Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. *Scientometrics, 86*(3), 687–703.

Yoon, J., Park, H., & Kim, K. (2013). Identifying technological competition trends for R&D planning using patent maps: SAO-based content analysis. *Scientometrics, 94*(1), 313–331.

Yusuf, Y., Gunasekaran, A., & Abthorpe, M. S. (2004). Enterprise information systems project implementation: a case study of ERP in Rolls-Royce. *Journal of Production Economics, 87*(3), 251–266.

Zippel-Schultz, B., & Schultz, C. (2011). Mediated and moderated effects of business and project planning on innovation projects in hospitals. *Creativity and Innovation Management, 20*(4), 296–310.