RJ10469 (A1006-004) June 16, 2010
Computer Science

IBM Research Report

Discovering Business Process Similarities: An Empirical Study with SAP Best Practice Business Processes

Anca Ivan, Rama Akkiraju
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099
USA

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

Discovering Business Process Similarities: An Empirical Study with SAP Best Practice Business Processes

Anca Ivan and Rama Akkiraju

IBM Almaden Research Center,
650 Harry Rd, San Jose, CA 95120

Abstract. Large organizations tend to have hundreds of business processes. Discovering and understanding the similarities among these business processes are useful to organizations for a number of reasons: (a) business processes can be managed and maintained more efficiently, (b) business processes can be reused in new or changed implementations, and (c) investment guidance on which aspects of business processes to improve can be obtained. In this empirical paper, we present the results of our study on over five thousand business processes obtained from SAP's standardized business process repository divided up into two groups: Industry-specific and Cross-industry. The results are encouraging. We found that 39% of cross-industry processes and 43% of SAP-industry processes are re-used across maps. Additionally, we found that 20% of all processes studied have at least 50% similarity with other processes. We use the notion of semantic similarity on process and process activity labels to determine similarity. These results indicate that there is enough similarity among business processes in organizations to take advantage of. While this is anecdotally stated, to our knowledge, this is the first attempt to empirically validate this hypothesis using real-world business processes of this size. We present the implications and future research directions on this topic and call for further empirical studies in this area.

Key words: business processes, process maps, discovery

1 Introduction

For the purposes of this paper, a business process means those 'structured activities or tasks' in an organization which, when executed in a specific way, 'produce a specific product or service for a specific customer' [14]. Examples of business processes include: accounts payable, accounts receivable, demand planning, order processing, employee payroll management, new-hire on boarding, sales promotion management, drug discovery management, clinical trial management etc. As can be noted in the example business process names above, some of them are applicable to most companies (cross-industry), while some are specific to some industries (industry-specific). For example, most companies have some kind of

employee payroll management, new hire on boarding, accounts payable and accounts receivable kind of processes. They are the cross-industry processes. Processes such as drug discovery and clinical trial management are processes that are specific to pharmaceutical industry. Demand planning business process is typically used in manufacturing oriented industries and sales promotion management is most typical in retail industry context. These are typically referred to as industry specific business processes.

A business process is said to be documented or modeled if the structure of the collection of process activities involved is represented and can be visualized as a model (eg: flowchart like diagram). A simplest form of such a representation could be a sequence of activities. More complex representations include forks, joins, parallel paths, decision nodes etc. A business process is said to be formally documented if its flowchart-like model adheres to formal rules of any one chosen business process representation language (eg: Business Process Modeling Notation BPMN [12], Unified Modeling Language [13], Petri nets).

Large organizations tend to have hundreds and sometimes even thousands of business processes. Organizations that are mature and disciplined in the way they run their business processes may maintain models of their business processes in repositories for reference and maintenance purposes. These repositories could be valuable assets for organizational learning. Analyzing these business process models and discovering and understanding the similarities among them can be useful to organizations in a number of ways. First, if similar aspects/tasks/activities of business processes are known business processes can be managed more efficiently. For example, if a software patch is to be applied to an application that supports a specific process activity, then knowing what all business processes use the same process activity is very helpful. This information helps in planning the business process unavailability and therefore eventually scheduling the business process unavailability optimally. Second, during new business process implementations or changes to existing business processes, those aspects of process activities that might be common with or similar to any of the existing process activities can be readily reused or leveraged for efficient implementations. Third, during mergers and acquisitions, knowing the similarities among the business processes of the two organizations involved can help with identifying opportunities for process standardization and consolidation. Finally, knowing similar process activities can also guide investment decisions if the identified similar process steps are associated with metrics that need further improvements. In summary, the need for understanding business process similarities within an organization is well-established.

Discovering business process similarities can be viewed from IT services provider perspective as well. Information Technology (IT) services companies are under constant pressure to deliver solutions quickly and cost effectively to their clients. One way to achieve this is by reusing assets developed for past clients after appropriate cleansing in the context of a new client project. Most IT services providers have some internal mechanisms in place to maintain assets from projects in some repositories. Business processes stored in these reposito-

ries could contain a wealth of knowledge and assets related to business process models, best practices, time taken to implement solutions etc. Leveraging the assets around these existing business process models can reduce project delivery times and improve project efficiencies. For example, some aspects of campaign management process in pharmaceutical industry might be similar to a trade promotions management process in retail industry. Therefore, the assets related to the campaign management processes such as best-practices, process definition documents, implementation guides, test scripts, and possibly even some code implemented for a specific client in pharmaceutical industry might be useful when implementing a trade promotions management process for another client in retail industry.

Thus far in this introduction, we have presented several arguments for why discovering and understanding similarities among business processes can be useful for organizations. However, an important question that we feel is not answered well in literature so far is: is there much similarity among business processes in organizations, in real-world or is it merely a topic of academic interest? We have not found any empirical studies presenting evidence one way or another. That is what motivated our work.

In this empirical study, we explore business process similarity with the aim of discovering similarities among a given set of business processes in an industry setting. We have analyzed over one thousand publicly available best practice business processes from SAP's business process repository to discover similarities among them within and across industries. The results are encouraging. We found that 20% of all processes studied have at least 0.5 similarity with other processes. We use the notion of semantic similarity on process and process activity labels to determine similarity. While this is a simple start, the results indicate that there is enough similarity among business processes in real-world organizations to take advantage of. While this is anecdotally observed thus far, to our knowledge, this is the first attempt to empirically prove this hypothesis using real business processes of this size. We present the implications and future research directions on this topic. The results can be applied and used by large organizations with many business processes and also for IT services companies that provide business process implementation services for their clients.

The rest of the paper is organized as follows. We discuss various aspects of business process similarities in Section 2 and discuss some of the techniques used in literature. We present the details of our experimental setup in Section 3. Section 4 presents our analysis of process similarities in SAP best practice processes. Finally, in Section 5 we discuss limitations of our study and call for additional empirical studies to substantiate the need for more research on this topic.

2 Background and Related work on Process Similarity

What information about business processes can be used to discover similarities? Business processes operate in specific contexts in organizations. Business processes have *names(labels), structure, semantics, and data flow.* They use *re-*

sources and manipulate resources(*system, people, and data*) and leave *traces* behind when executed. The efficiency and effectiveness of business processes can be measured by *metrics*. Business processes serve specific *business objectives* that can be measured by *key performance indicators*. Business process designers and analysts document information about processes and process steps in plain text in *design documents*. Sometimes there might be rich text (attachments, diagrams with annotations etc) about processes in process design documents. Business processes can be *classified* along many dimensions: *industry, scenario groups, scenarios, functional areas, organizations (that are responsible for managing and maintaining them), user groups and roles (that use the processes)*. Business processes have *dependencies* on other processes and are *related* to other business process. When business processes are not functioning normally or as desired, certain aspects of business will be at *risk*. Business processes need to be maintained and updated and when *outages (planned or unplanned)* occur, there will be *business impact*. All of this information about processes can be used for determining process similarity.

Obtaining all of the information about business processes at once is hard. The organizational structure of companies makes it difficult to go to one single source to obtain comprehensive information about business processes. In the past, there weren't enough mature tools that supported formal business process modeling, analysis, and simulation. Even if they existed, they were considered academic and didn't receive much adoption in the industry. So, not many companies formally documented their business processes. Recent emergence of industry standards such as Business Process Modeling Notation [12] and Business Process Execution Language [15] combined with the maturity and accessibility of vendor tools make the goal of discovering business process similarities more achievable than it was in the past.

Our related work analysis on business process similarity matching is not comprehensive since we are concerned with presenting the results of our empirical study. We note that much of the past work done in business process similarity matching is done based on matching of process labels, process structure (control flow) and process execution semantics. Dumas et al. [4] present a nice summary of various techniques that have been applied to conduct process similarity matching in their paper. They note the usage of string-edit distance based approaches for label matching, graph matching techniques [1–3] for process control flow matching and process mining techniques for matching execution semantics (traces, logs) [5, 6]. Simulation and causal footprint analysis [7] have also been used for matching. Much of this work focuses on matching pairs of business processes. However, in our work, we are concerned with matching a query with a repository of business processes.

Clustering and machine learning communities have looked at repositories of business processes. For example, Lee H.S [8] generates hierarchical clusters from a set of business processes using the notions of cohesion and coupling. The author uses clustering as a means to find out related and dependent processes. This work does not directly focus on finding a set of matching processes (from a repository

of processes) with the purpose of reuse in mind. J. Melcher and Sees [9] applied clustering to SAP reference models using process metrics values for finding (structurally) similar processes among business process collections. The process metric values of various processes are compared to obtain a heatmap. This visual technique is used for clustering. This is one empirical work that we are aware of on SAP reference models. We perform semantic matching using process and process step labels whereas this work used process metrics information. In another work Jung and Bae et al [10] apply hierarchical machine learning to discover process similarities among a group of processes. In their work they first transform business process models to vector models based on their structures such as activities and transitions, and the vectors are compared by Cosine similarity measure. Finally, the models are clustered by the agglomerative hierarchical clustering algorithm. B. Srivastava [11] uses process features to derive summaries on groups of processes. These latter two works aim to address the same problem of leveraging existing processes for future process implementations.

Although, it not the main focus of this empirical paper, it is interesting to note that there is no body of work leveraging multiple attributes of business processes at the same time - perhaps because of the difficulty of obtaining that data. But it is increasingly becoming possible to obtain that information. We believe that a combination of text analytics, clustering and structure matching that takes into account data, resource flows into account will yield more accurate and precise matches. This is the subject of our next study. For now, we turn our attention to our study.

3 Data Facts

We believe that there is enough similarity among business processes within a company and within and across industries that discovering and understanding these similarities is beneficial for individual organizations as well as IT services providers. We want to test this hypothesis with our experiments.

Process complexity. The data analysis was performed on 21 solution best practice maps from the SAP Solution Composer tool [16]. Solution Composer keeps these maps as XML files. Unfortunately, the XML files have no consistent schema. Our first challenge was extracting the processes and their structure for every map. Once the data was extracted and saved a consistent computer readable form, we ran several experiments to determine the nature of the data. Our conclusions are presented next.

On average, the solution maps have 160 processes. The cross-industry maps have 1916 processes, the SAP industry maps have 3383 processes (total of 5299 processes). Table 1 shows the number of processes defined for a sample set of cross-industry and industry maps.

Each process has an average 6 or 7 process steps; Figure 1 shows the distribution of processes as a function of number of steps contained in the process. A very small number of processes have more than 20 steps: for example, the *Campaign Planning and Execution in CRM* process from the *SAP Service and*

Cross Industry Map	No.	Industry Map	No.
mySAP Product Lifecycle Management	47	Aerospace	192
Channel Management	377	Banking	244
SAP Business One integration for SAP NetWeaver	19	Defense	149
SAP Radio Frequency Identification (RFID)	62	Defense Logistics	220
SAP Business One 2005	239	Mining	236
Field Applications	344	Public Sector	177
E-Commerce	256	Higher Education	171
SAP Business One 2004	239	Hospitality	272
SAP CRM Powered by SAP NetWeaver	208	Insurance	113
SAP NetWeaver	82	Research	106
SAP Global Trade Services	43	Utilities Retail	239

Table 1. Number of processes for a sample set of industry and cross-industry maps.

Asset Management cross-industry map has 27 steps, and 13 processes from industry maps including *Management of Internal Controls* from *Mining* industry map which has 56 steps, and *Local Close* from *Banking* industry map which has 37 steps).

Process duplication. Given the large number of processes defined by both industry and cross-industry maps, the next question is whether these processes are unique to their respective domains. Figure 2 highlights the process duplication across maps. In this experiment, two processes are considered to be duplicates if their names are identical. The Y axis represents the number of processes that belong to 1, 2, 3 or more industries. In the case of cross-industry maps, 648 processes are unique to a single map, 295 processes can be found in 2 maps, 140 can be found in 3 maps, and 65 processes can be found in more than 3 maps. For example, the *Self-Service Support through FAQ and Solution Search* process belongs to the following cross-industry maps: *SAP Service and Asset Management*, *E-Commerce*, and *Channel Management*. In the case of cross industry maps, 39% of process are duplicated across maps; in the case of SAP-industry maps, 43% are duplicated across maps.

4 Data Analysis

In order to evaluate the process similarity, we are using 3 approaches: (1) identify process duplication (as described above), (2) identify common steps amongst processes, and (3) compute the semantic similarity score for the process names.

The semantic similarity score is given by the following equation:

$$S = 2 * n_{CS}/(n_P + n_Q)$$

where n_P and n_Q are the numbers of steps contained in the structures of processes P and Q, and n_{CS} is the number of common steps.

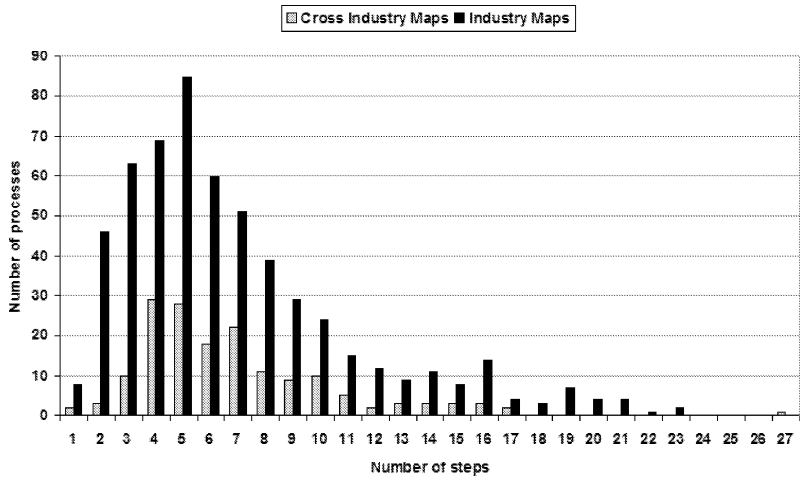


Fig. 1. Process complexity as a function of number of steps.

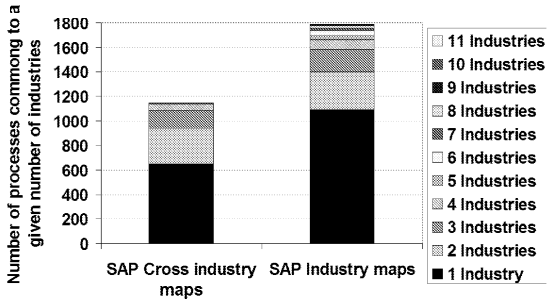


Fig. 2. Process duplication across industry and cross-industry maps.

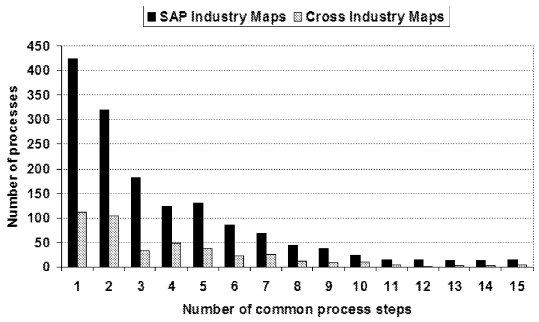


Fig. 3. Process commonality across industries as a function of number of steps.

Step-based similarity. Figure 3 shows how often business processes share process steps. The X axis represents the number of process steps shared by two processes. The Y axis represents the number of processes that share the given number of process steps. In the SAP industry maps data set, the *Complaints Processing with CRM Mobile Service for Handheld* process from the *Logistics Service Providers* map and the *Service Order Processing with CRM Mobile Service for Handheld* process from *Utilities* map have 3 steps in common (similarity score of 0.5): (1) Synchronize data, (2) System replicates data, and (3) Assign business partner and contact person.

Process similarity scores. Figure 4 shows the process similarity trends for both data sets; we computed the similarity scores between all pairs of processes defined in each domain (cross-industry, and industry-specific). The similarity scores are high for a small number of processes, and then they drop. Only about 20% of processes have similarity scores higher than 0.5: for example, in the cross-industry maps, the similarity score between *Quotation Processing with CRM Mobile Sales* and *Activity Processing with CRM Mobile Sales* is 0.8. This means that 80% of the processes either have no variant or the variant is so different in name that semantic matching is not discovering the variant.

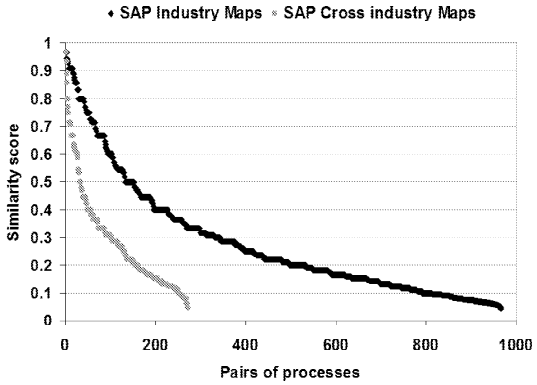
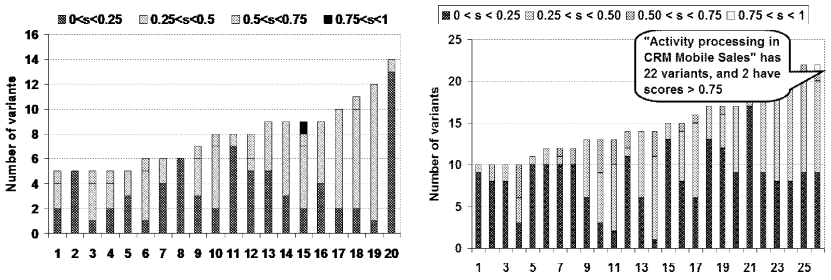


Fig. 4. Similarity scores.

Process variants. In order to find process variants, we used the semantic similarity algorithm previously described for our experiments. Both graphs show only a subset of the processes that have variants. For presentation purposes, we only show a subset of the process variants that have representative variants (many of the processes have very low similarity scores). In the case of cross-industry maps, most processes are not connected (similarity scores are very low). This makes sense because an enterprise runs only one instance of a process.

There are only a few processes that have variants and were found by the semantic similarity engine; for example, "Complaints Processing with CRM Mobile Service" and "Complaints Processing with CRM Mobile Service for Handheld". In the case of industry maps, the similarity scores are higher because there are more processes related across multiple industries.



(a) Cross-industry maps. (b) Industry maps.

Fig. 5. Process variants.

The results indicate the following. (1) There is enough similarity among these processes studied. 39% of cross-industry processes and 43% of SAP-industry processes are re-used across maps. Additionally, we found that 20% of all processes studied have at least 50% similarity with other processes. (2) The amount of similarity among processes in cross-industry segment is a little bit smaller than (39%) that discovered among processes in industry segment (43%). While it would be too early to draw any conclusions based on these experiments which only consider the process and process step labels as similarity metrics, the results reveal opportunities for some further areas of study. This result when substantiated further might mean that there are fewer opportunities to leverage similarities among business processes within an organization than among best practice processes developed for many industries by IT services provider. IT services providers might benefit more from conducting business process similarity matching than individual organizations. This result makes one assumption that individual organizations all have non-duplicating common best practices. We know that that is seldom the case. Often companies end up with duplicate processes and systems due to mergers and acquisitions. So, this result does not apply to that case since the data assumes that there are standardized processes within an organization. In any case, this calls for further empirical studies with business process data within individual organizations as well as with data of best practices from IT services providers to really understand the nature of similarities in business processes within companies, within industries and across industries.

5 Discussions and Conclusions

The data used for analysis in this work is from the best practice process repository of a business process implementation software vendor, SAP. Individual companies implement business processes that pertain to their industry. They don't implement all processes for all industries. Therefore, the results from analyzing industry-specific processes apply more to IT services providers' that implement business process software for companies than to individual companies. The cross-industry business processes apply more generally to companies in any industry. So, these results apply to both individual companies and IT services providers. Our study, presently, uses semantic similarity of process and process activity labels to determine similarities among processes. One might argue that semantic matching of labels alone is a weak indication of similarity if the structure is not considered. While we acknowledge that in general, we argue that for the specific data set we considered that is not an issue. This is because SAP business process repository uses vocabulary that is standardized across all of SAP's software. For example, the word 'sales order' in a process activity would mean the same when it is found in any other process activity in any business process within SAP domain. Therefore, label matching works well in this domain. We acknowledge the limitations of this study both in the amount of information about business processes that was used and in the simplicity of technical approaches used to discover similarities among business processes. The main reason for limiting the study to these aspects is the lack of publicly available data on some of the other aspects of business processes (such as control flow, data flow, resource flow etc) for the data we considered. The results presented in this paper can be treated as a call for further empirical studies to discover similarities among business processes in an industry setting. As of the time of writing of this paper, we were fortunate to get access to a large repository of internal business processes of a large company (one of the 30 Dow Jones companies who guide the Dow Jones Industrial Average Index) with detailed control flow, data flow, resource flow and detailed design documentation. We are currently studying that data and are exploring the application of machine learning, clustering and a combination of graph matching, text analysis techniques and semantic matching algorithms to determine process similarities.

References

1. R. Dijkman, M. Dumas, and L. Garcia-Banuelos: Graph matching algorithms for business process model similarity search. Proceedings of BPM 2009, Ulm, Germany (2009)
2. S. Melnik, H. Garcia-Molina, and E. Rahm: Similarity flooding: A versatile graph matching algorithm (extended technical report). Technical Report 2001-25, Stanford InfoLab (2001)
3. C. Li, M. U. Reichert, and A. Wombacher: On measuring process model similarity based on high-level change operations. Technical Report TR-CTIT-07-89, CTIT, Enschede, The Netherlands (2007)

4. Dumas M., Garcia-Banuelos L., Dijkman R: Similarity Search of Business Process Models. http://sites.computer.org/debull/A09sept/marlon.pdf
5. Aalst, W. van der, Weijters, A., and Maruster, L: Workflow Mining: Discovering Process Models from Event Logs. IEEE Transactions on Knowledge and Data Engineering, 16 (9), 1128-1142 (2004)
6. Agrawal, R., Gunopulos, D., and Leymann, F.: Mining Process Models from Workflow Logs. Sixth international conference on extending database technology, pp. 469-483 (1998)
7. B. F. van Dongen, R. M. Dijkman, and J. Mendling: Measuring similarity between business process models, Proc. of CAiSE, volume 5074 of LNCS, pages 450-464. Springer (2008)
8. Lee H.S.: Automatic clustering of business processes, Business systems planning European Journal of Operational Research Volume 114, Issue 2, Pages 354-362 (1999)
9. Joachim Melcher, Detlef Seese: Visualization and Clustering of Business Process Collections Based on Process Metric Values, 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp.572-575 (2008)
10. J.Y. Jung, J. Bae, and L. Liu: Hierarchical Clustering of Business Process Models, International Journal of Innovative Computing, Information and Control, Volume 5, Number 12 (2009)
11. B. Srivastava: Summarizing Business Processes; Manuscript (2009)
12. Business Process Model and Notation, http://www.omg.org/spec/BPMN/
13. Unified Modeling Language, http://www.omg.org/technology/documents/formal/uml.htm
14. Wikipedia, http://www.wikipedia.org/
15. Web Services Business Process Execution Language, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-specification-draft.html
16. SAP BUSINESS MAPS, SOLUTION COMPOSER, http://www.sap.com/solutions/businessmaps/composer/index.epx