Key Deficiencies of Semantic Business Process Search

Avi Wasser and Maya Lincoln

University of Haifa, Israel
awasser@haifa.ac.il, maya.lincoln@processgene.com

Abstract In recent years, researchers have become increasingly interested in developing frameworks and tools for searching business process model repositories. While research on searching structured repositories has been extensive, little attention was dedicated to searching business process content within unstructured repositories, such as the Web. We demonstrate why current search technologies are not useful for extracting process content from the Web, and explain the core reasons for the deficiency. We then express the requirements for a framework that could overcome the presented shortcomings.

Keywords: Business process search, Business process repositories, Natural language processing, semantic search, operational search.

1 Introduction

Business Process Models (BPMs) are considered an important mine of organizational knowledge, and therefore are a major source for searching and retrieving operational and enterprise related data [18].

Researchers have become increasingly interested in developing methods and tools for retrieving information from business process repositories [3,14,19]. While research on searching structured repositories has been extensive, little or no attention was dedicated to searching business process content within unstructured repositories, such as the Web. Such repositories are constantly becoming more extensive, and are accessible to a wide user population through search engines.

Two common methods for retrieving information from a repository are querying and searching. Querying is aimed at retrieving information using a structured query language. The significance of querying business processes has been acknowledged by BPMI[1] that launched a Business Process Query Language (BPQL) initiative. Searching, on the other hand, allows information retrieval using keywords or natural language and was shown to be an effective method for non-experts.

Research in the field of business process retrieval has mainly focused on *semantics* and *structural* similarity analysis techniques [3,15,4,11]. Using these frameworks one can retrieve process models that either contain semantically

[1] Business Process Management Initiative, http://www.bpmi.org/

R. Meersman et al. (Eds.): OTM 2014 Workshops, LNCS 8842, pp. 339–346, 2014.

related components (e.g. activity names with a specified keyword) or match a requested graph structure: e.g. that presents a sequence of activities. While these methods can be applied on structured process repositories, it is practically impossible to apply them on the unstructured Web.

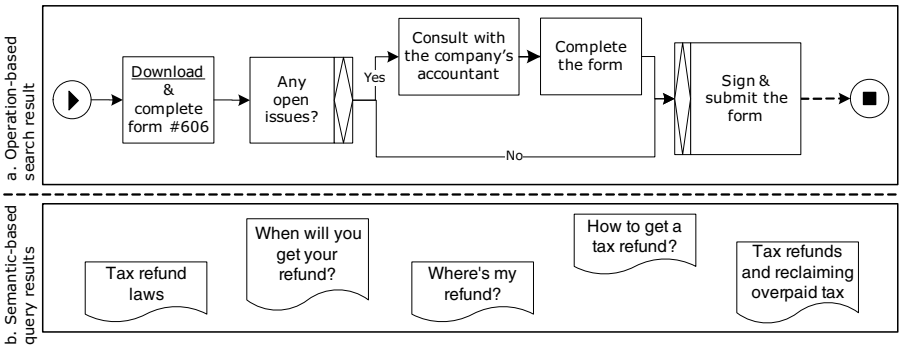

Fig. 1. An example of search results for "how to claim a tax refund"

In order to illustrate why semantic search is not adequate for process retrieval from unstructured repositories, we will present a motivating example, as follows. Consider an employee interested in finding out "how to claim a tax refund." An expected outcome of this retrieval request would be a *process model* that represents the order of activities that one should follow in order to achieve the required process goal, as illustrated in Fig 1a. The benefit of such a retrieval framework is that the result is ready for execution. Without any preliminary knowledge of the underlying repository structure, the user can receive a full-fledged process model.

The retrieval output is related to the search phrase in *operational* terms. For example, Fig 1a provides a segment that is *not* similar semantically to the search phrase text. Specifically, all three search phrase terms ("Claim", "Tax" and "Refund") are not represented by any of its activities. Such "how-to" questions are hard to fulfill using common query languages due to the complex logic that is embedded within such questions [4] and especially without specific knowledge on process structure and activity naming. Therefore, using querying techniques on the Web, would yield a list of data items (e.g. Web pages, or media items) with semantically similar titles, as illustrated in Fig 1b. Such outcome does not tell the user "how-to" fulfill the process goal in a structured and operational manner.

In this work we present the key deficiencies of semantic business process search. We demonstrate the current shortcomings using examples from top four search engines.

The rest of the paper is organized as follows: we present related work in Section 2, positioning our work with respect to previous research. In Section 3 we present the major shortcomings of current Web search engines in extracting process data. We discuss future research elaborations and conclude in Section 4.

2 Related Work

Related works include query and search techniques in BPM. Works such as [16,17,4,2,7] query business process repositories to extract process model (graph) segments. Such methods require prior knowledge of the structure of the process repository and the exact notation that is used to express it. Therefore, they are not adequate for search on the Web that should work well even without prior knowledge regarding the process repository.

Keyword search on general tree or graph data structures can also be applied to process repositories [10,8,9]. These methods allow users to find information without having to learn a complex query language or getting prior knowledge of the process structure. Therefore, this method is also applied by leading business process management (BPM) software vendors, such as ProcessGene[2], SAP[3], Oracle[4] and others. Some works extend the tree and graph keyword search methods to support a more intuitive interface for the user by enabling searches based on natural language [13,12]. According to [1], the straightforwardness of a natural language makes it the most desirable database query interface. The retrieved information in both keyword and natural language search methods is in the form of single process model components such as activities and roles that are semantically similar to the searched phrase. These techniques are merely relevant to process search on the Web, since in this case (a) users are seeking to receive a complete process; and (b) the expected process result is usually not related *semantically* to the search phrase, but rather *operationally*.

The work in [14] extends the above line of works. This work supports the retrieval of complete process segments by applying dynamic segmentation of the process repository. The search result is a compendium of data (a segment of a business process model) related to the operational meaning of the searched text. Nevertheless, as all other works, this method relies also on a process-structured database, and cannot work "as is" on an unstructured repository, such as the Web.

Another line of work focuses on automatic construction of process data ontologies. The work in [5] proposes a query-by-example approach that relies on ontological description of business processes, activities, and their relationships, which can automatically be built from the workflow models themselves. The work in [6] automatically extracts the semantics from searched conceptual models, without requiring manual meta-data annotation, while basing its method on a model-independent framework. The work in [14] automatically extracts and uses the operational layer (the "how-to") and the business rules encapsulated in a process repository. Such automatic ontology extraction techniques are important for analyzing data encapsulated in the Web. Nevertheless, the current research literature is based solely on process-flow structured repositories and not on unstructured repositories such as the Web.

[2] `http://www.processgene.com`
[3] `http://www.sap.com`
[4] `http://www.oracle.com`

3 Key Deficiencies of Semantic Business Process Search

Current Web searches are based on keyword queries and semantic similarity lookups. This makes data extraction relatively easy and simple for users. Nevertheless, and as demonstrated in Section 1, it is practically impossible to extract processes from the Web using current semantic search engine technology. This is an inherent material weakness that in our opinion presents a significant barrier for the evolution of Web usability.

The main search engines (e.g. Google[5], Microsoft Bing[6], Yahoo[7], Ask[8] and others) are still at experimental, initial phases of enabling Web process-searches. For example, recent R&D efforts of Google yield lists of "how to" instructions for very limited process sets. For instance, when searching in Google "how to issue an invoice," a set of related documents and media is retrieved, without any process-formatted results, as illustrated in Fig. 2. However, in some cases, we identified initial attempts to retrieve instruction-based results for certain "how to" queries. These results are presented before the standard Google search results, within a dedicated frame, in a list-format, which is a first step in aiming to retrieve and present process-flow formats. This presentation is still at a preliminary phase as Google requests users' feedback regarding the quality of these instruction lists. An example of such a list resulted from the search phrase: "How to take a screenshot from Windows" is presented in Fig. 3.

Besides these attempts to present somewhat process-driven results, we note that standard search results for "how-to" or "process-driven" queries are very limited, again due to the aforementioned material weakness of semantic search. We sampled the top 4 search engines (Google, Bing, Yahoo and Ask) with the following examples of process-search scenarios, as presented in Table 1. As demonstrated, current search technologies cannot provide adequate results. We included not only business process but also personal process queries as we believe that the size and amount of data presented in the Web will also extend the scope of process related searches beyond the domain of BPM.

Clearly, these results encompass a large, unstructured, set of data with a low level of usability:

- Results are not presented in standardized process notations - nor in basic flowchart formations. Therefore, practically, for the end-user, it is not possible to deduct an actual process from search results.
- It is not clear what the required steps are and what is the order of activities for achieving the process goal.
- It is also not possible to assess the quality and relevance of the suggested results from an operational viewpoint - as ranking is based on semantics and not on operational characteristics of a process.

[5] `https://www.google.com`
[6] `http://www.bing.com`
[7] `https://www.yahoo.com`
[8] `http://www.ask.com`

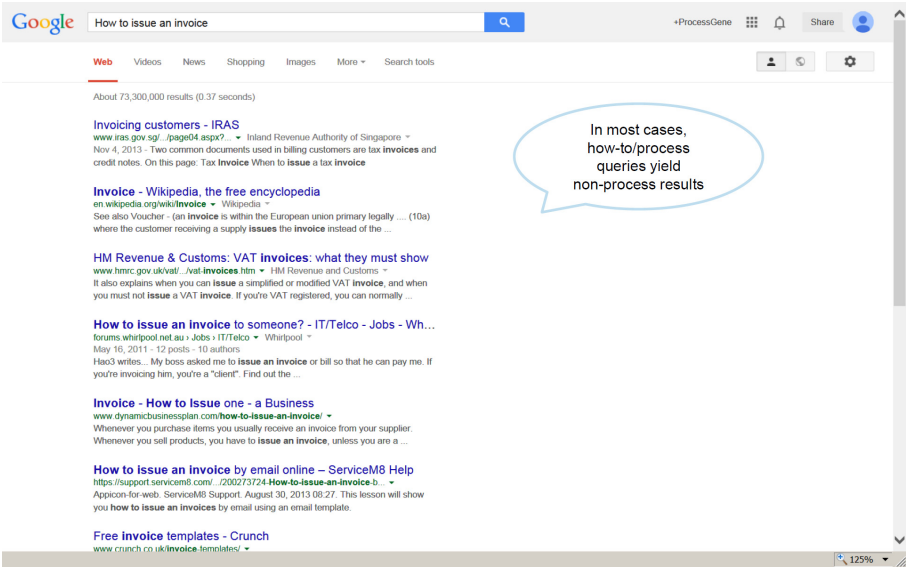

Fig. 2. An example of search results for "how to issue an invoice"

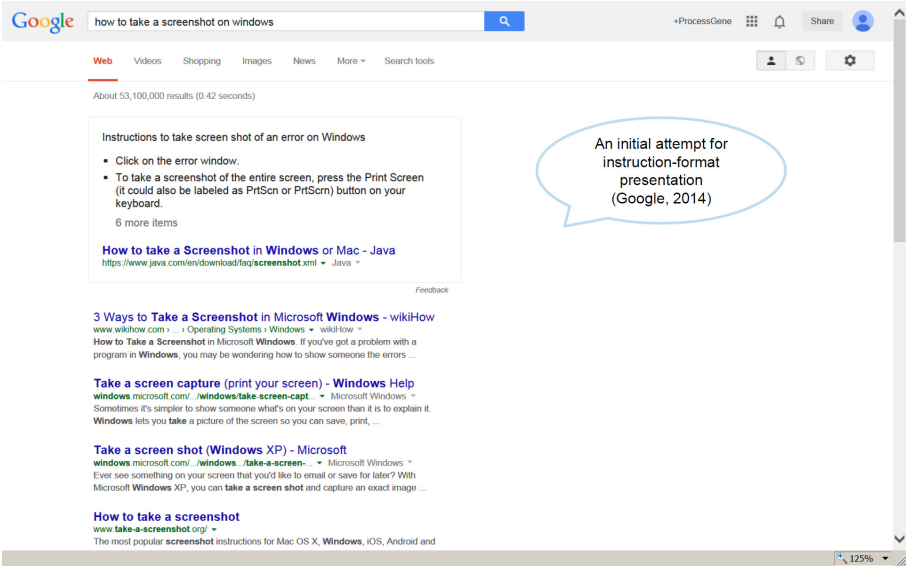

Fig. 3. An example of list-formatted search results for "how to take a screenshot on Windows"

Table 1. Examples of process-search scenarios using current Web search technologies

Query	Result Samples
"Get a student loan in London"	Top of the page: Ads on general loans. Several non relevant ads. Thousands of loosely related pages.
"Monitor first pregnancy in Baltimore"	Top of the page: Ads on abortion and monitor sales. highly ranked results- damages of using cocaine during pregnancy, Craig-list results on pregnancy issues. Thousands of loosely related pages.
"Select and qualify suppliers"	Mainly pages of HR / personnel recruitment companies.
"Arrange a wedding"	Tripadvisor recommendations on wedding locations, ads on churches and, explanations on flower bouquets.

Hence, these samples along with the elaborated example presented in section 1 demonstrate current process-search deficiencies and the need for an alternative framework that will support such Web searches.

Finally, in order to estimate the demand for such a solution we ran targeted Web queries using Google's keyword planner[9] and Bing's Keyword Research tool[10]. These tools aim to monitor, on an ongoing basis, the frequency in which any certain keyword (or an ordered group of keywords) is submitted. It turns out that "how-to" related keywords are searched over 5.5 million times every month by Google and Bing together, as demonstrated in Table 2. As of July 2014 Google and Bing hold over 73% of the search engine market[11]. Therefore, an extrapolation of the above results to the rest of the search engines brings us to over 7.6 million "how-to" related searches per month, and over 91.4 million such searches per year.

Table 2. Number of "how-to" related searches per month (August, 2014)

Keyword	Google	Bing	**Total**
"how to"	450,000	4,683,152	5,133,152
"steps for"	170	0	170
"process"	135,000	0	135,000
"procedure"	74,000	75,449	149,449
"list of activities"	1,600	80	1,680
"checklist"	74,000	0	74,000
"flowchart"	60,500	2,997	63,497
"process flow"	4,400	0	4,400
Total	799,670	4,761,678	**5,561,348**

[9] `https://adwords.google.com/KeywordPlanner`

[10] `https://www.bing.com/webmaster/diagnostics/keyword/research`

[11] According to Netmarketshare, `http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0`

Apparently, there is a relatively high demand for such queries, and despite this large target market there is no feasible solution for conducting these process-driven searches.

4 Conclusions

In this work we presented the key deficiencies of semantic business process search within the Web. The need for unstructured search capabilities exists not only for organizations, but also for a large audience of individuals that seek an accessible solution for "how to" queries. As future work, we propose to structure a framework that could overcome the shortcomings of existing search technologies within unstructured repositories. Such a framework should describe: (1) how to extract full-fledged process models out of unstructured repositories; (2) a process-based ranking and relaxation mechanisms. In addition, it will be required to provide an applicative case study and experiments to measure the efficiency of the proposed framework. It is hoped that by expanding search and query capabilities of processes within the Web, users will be able to extract operational knowledge more simply and efficiently.

References

1. Androutsopoulos, I., Ritchie, G.D., Thanisch, P.: Natural language interfaces to databases-an introduction. Natural Language Engineering 1(01), 29–81 (1995)
2. Awad, A.: BPMN-Q: A Language to Query Business Processes. In: EMISA, vol. 119, pp. 115–128 (2007)
3. Awad, A., Polyvyanyy, A., Weske, M.: Semantic querying of business process models. In: 12th International IEEE Enterprise Distributed Object Computing Conference, pp. 85–94. IEEE (2008)
4. Beeri, C., Eyal, A., Kamenkovich, S., Milo, T.: Querying business processes with BP-QL. Information Systems 33(6), 477–507 (2008)
5. Belhajjame, K., Brambilla, M.: Ontology-based description and discovery of business processes. Enterprise, Business-Process and Information Systems Modeling, 85–98 (2009)
6. Bozzon, A., Brambilla, M., Fraternali, P.: Searching repositories of web application models. In: Benatallah, B., Casati, F., Kappel, G., Rossi, G. (eds.) ICWE 2010. LNCS, vol. 6189, pp. 1–15. Springer, Heidelberg (2010)
7. Goderis, A., Li, P., Goble, C.: Workflow discovery: the problem, a case study from e-Science and a graph-based solution (2006)
8. Guo, L., Shao, F., Botev, C., Shanmugasundaram, J.: XRANK: Ranked keyword search over XML documents. In: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 16–27. ACM (2003)
9. He, H., Wang, H., Yang, J., Yu, P.S.: BLINKS: ranked keyword searches on graphs. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 305–316. ACM (2007)
10. Hristidis, V., Papakonstantinou, Y., Balmin, A.: Keyword proximity search on XML graphs (2003)

11. Karni, R., Wasser, A., Lincoln, M.: Content analysis of business processes. International Journal of E-Business Development (2014)
12. Katz, B., Lin, J., Quan, D.: Natural language annotations for the SemanticWeb. On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE, 1317–1331 (2010)
13. Li, Y., Yang, H., Jagadish, H.V.: NaLIX: A generic natural language search environment for XML data. ACM Transactions on Database Systems (TODS) 32(4), 30 (2007)
14. Lincoln, M., Gal, A.: Searching business process repositories using operational similarity. On the Move to Meaningful Internet Systems: OTM, 2–19 (2011)
15. Markovic, I., Pereira, A.C., Stojanovic, N.: A framework for querying in business process modelling. In: Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI), Munchen, Germany (2008)
16. Momotko, M., Subieta, K.: Process query language: A way to make workflow processes more flexible. In: Benczúr, A.A., Demetrovics, J., Gottlob, G. (eds.) ADBIS 2004. LNCS, vol. 3255, pp. 306–321. Springer, Heidelberg (2004)
17. Shao, Q., Sun, P., Chen, Y.: WISE: a workflow information search engine. In: IEEE 25th International Conference on ICDE 2009, pp. 1491–1494. IEEE (2009)
18. Wasser, A., Lincoln, M., Karni, R.: Accelerated enterprise process modeling through a formalized functional typology. In: van der Aalst, W.M.P., Benatallah, B., Casati, F., Curbera, F. (eds.) BPM 2005. LNCS, vol. 3649, pp. 446–451. Springer, Heidelberg (2005)
19. Wasser, A., Lincoln, M., Karni, R.: Processgene querya tool for querying the content layer of business process models. In: Demo Session of the 4th International Conference on Business Process Management, p. 18 (2006)