Collaborative Evaluation to Build Closed Repositories on Business Process Models

Hugo Ordoñez[1], Juan Carlos Corrales[1], Carlos Cobos[2], Leandro Krug Wives[3] and Lucineia Thom[3]
[1]*Telematics Engineering department, University of Cauca, Sector Tulcán, Popayán, Colombia*
[2]*Systems engineering department, University of Cauca, Sector Tulcán, Popayán, Colombia*
[3]*Institute of Informatics, Federal University of Rio Grande do Sul, Caixa Postal 15.064, Porto Alegre, RS, Brazil*

Keywords: Business Process Relevance, Business Processes Management, Collaborative Methodology, Business Process Search Evaluation.

Abstract: Nowadays, many companies define, model and use business processes (BP) for several tasks. BP management has become an important research area and researchers have focused their attention on the development of mechanisms for searching BP models on repositories. Despite the positive results of the current mechanisms, there is no defined collaborative methodology to create a closed repository evaluation for these search mechanisms. This kind of repository contains some closed BP predefined lists representing queries and ideal answers to these queries with the most relevant BPs based on a set of evaluation metrics. This paper describes a methodology for creating such repositories. To apply the proposed methodology, we built a Web tool that allows to a set of evaluators to make relevance judgments in a collaborative way for each one of the items returned according to predefined queries. The evaluation metrics used can measure the consensus degree in the results, therefore confirming the methodology feasibility to create an open access, scalable and expandable closed BP repository with new BP models that can be reusable in future research.

1 INTRODUCTION

Currently, many companies define, model, and use business processes (BP) for several tasks such as manufacturing, services, purchasing, inventory management and others. With the advances in technology development, the impact of BP management has become an increasingly important research area in academic and business fields. As a result, big effort has been dedicated to the development of mechanisms to search and discover reusable components (Škrinjar and Trkman 2012) for defining new BP adjustable to current requirements of the organization. These efforts are aimed at providing companies a starting point to improve their trading activities.

Therefore, these mechanisms should be evaluated to find their inconsistencies, fix them and ensure the proper implementation of their functional purpose. Besides, there is still a lack of closed repositories in business process evaluation that would allow to compare the performance of two or more BP searching techniques in the same conditions. This also could help to find the shortcomings and to make improvements to these techniques.

This paper presents a collaborative evaluation methodology to build closed repositories. It also presents and discusses the outcomes obtained after applying the proposed methodology. To this end, we have developed and used a tool that implements this methodology and uses a BP searching mechanism to return a smart BPs list created with the BPs to be evaluated on each query. Thus, evaluators do not have to evaluate all existing BPs within the repository.

The methodology is proposed to build closed repositories' evaluation while taking into account the opinion of an expert group from a collaborative perspective. In this sense, each expert makes relevance judgments between BPs reported as results by a searching mechanism and a BP defined as query. Then the BP query mechanisms can use the repository to evaluate the quality in their searching process.

This paper presents two specific contributions: first, an evaluation methodology to create closed repositories of BPs taking into account the opinions

Ordoñez H., Carlos Corrales J., Cobos C., Krug Wives L. and Thom L..
Collaborative Evaluation to Build Closed Repositories on Business Process Models.
DOI: 10.5220/0004881203110318
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 311-318
ISBN: 978-989-758-029-1

of a group of experts; and, second, an open access BP repository (motivated by the approach proposed in(Kunze and Weske 2012)) with a hundred BP models from the telecommunications and geo-referencing domain.

The rest of the paper is organized as follows: Section 2 describes related work and evaluation methodologies for BP model searching mechanisms. Section 3 presents the proposed methodology for collaborative assessment. Section 4 describes a Web tool specially developed to allow the projected methodology's application. Section 5 describes the repository. Section 6 describes a case study, and Section 7 presents the conclusions and future works that are expected in the short term.

2 RELATED WORK

Despite the progress in the development of tools for searching and discovering BPs (Rosa, Arthur et al. 2010; Kunze 2013), to date there are no formal methodologies to evaluate these mechanisms.

Regarding the above, some related works propose evaluation methodologies and experimental setups centered on the evaluation of tools for discovering Semantic Web Services (SWS).

Consequently, these experimental setups can serve as a starting point to create a formal evaluation methodology for the results reported by BP searching tools.

2.1 Evaluation on BP Searching

Regarding the BP searching task, some metrics have been defined to measure or evaluate the degree of precision and relevance of the results reported by proposals for finding similarities between BPs (Dijkman et al., 2011); (Becker and Laue, 2012). Among those proposals are: linguistic, focused on the name or description of each BP element (Koschmider et al., 2011); association rules, focused on the historical execution of BP tasks which are recorded in log files; and genetic algorithms that integrate more data as inputs, outputs, edges, and nodes in the search process (Turner, 2010). In addition to these proposals, there are further approaches centered on searching BP models within repositories using proprietary languages or methods for excecuting queries (La Rosa et al., 2011); (Yan et al., 2012)

2.2 Evaluation Methodologies

In (Tsetsos et al., 2006), for instance, an evaluation system for Semantic Web Services (SWS) discovery based on information retrieval (IR) theories is proposed. there two similarity schemes are evaluated: 1) A Boolean schema that sets two values, 0 or 1 for similarity degrees, and a correspondence between a query service and a comparison service, where "1" means that two services have some level of affinity, and "0" when they have no affinity; 2) A scale of similarity values (i.e., numerical values in the range [0-1], corresponding to fuzzy terms like "relevant", "irrelevant", and so on) that allows us to sort the results according to similarity levels, which present the query services and a comparison service. In this case, the evaluation is made according to the equivalence between the services sorted by the experts and the result obtained by the tool.

In (Küster and König-Ries, 2009)a services collection is shown. This collection contains three different evaluation scales that were used to classify the relevance of the reported results in a query. They have used three schemes: 1) A binary one, which has been most commonly used, where “1” determines that there is a degree of relevance and “0” that there is no relevance at all; 2) One-dimensional graded relevance that is a multi-valued scale to measure the similarity between two services; 3) A Multi-dimensional graduate importance, which provides a multi-scale to evaluate different aspects (equivalence, scope and interface, among others) between two services.

Moreover, (Dijkman et al., 2011)state that there is a considerable research gap for comparing different approaches for searching BPs because the evaluation process has only been based on similarity metrics evaluation, and therefore it is interesting to evaluate several of these approaches in the same scenario or closed repository.

As noted in previous works, so far there is no method or methodology for BP evaluation that integrates several experts to collaboratively build closed repositories of BPs that could serve as a basis for evaluations involving semantics and structure on BP searching.

Considering the description above, in (Kunze and Weske, 2012)an open library available to all community members is proposed. This library shares the BP's information and repositories following a few guidelines. For this reason, it is important to contribute to the definition of a BP repository based on the ideas expressed in: A successful BP repository depends on having a good searching engine allowing the retrieval of the desired process models in a short time period. In addition, due to the evaluations made on the repository, it may act as a closed document collection where, for each proposed query, the resulting BPs and their corresponding relevance levels are known.

3 EVALUATION METHODOLOGY

The proposed collaborative evaluation methodology is divided into three stages: individual evaluation, searching for consensus on discordant evaluations, and results refinement. The methodology arises as a consolidation instrument which allows a set of judges to make judgments in relation to relevant results against a BP query in a collection (or list) of BP previously stored.

Indeed, the results considered relevant by the panel of judges will be those that represent the ideal responses for each query in the closed repository built.

The evaluation takes a set of BPs from the repository, defined as $Q=\{bp_1,bp_2,bp_3...bp_n\}$, which represents each of the queries. For each query, a resulting list of items T is evaluated, where $T<=M$ (in order to decrease the workload of judges), and M is all the BP existent in the repository. Each item of the resulting list is evaluated using a Likert scale containing the following concepts: very relevant, relevant, quite relevant, not very relevant, and irrelevant. This scale is defined because two BPs may have different similarity levels in relation to each other. The weight (w) assigned to each concept of relevance is w={1, 0.75, 0.50, 0.25, 0} in the scale and, therefore, the overall relevance level (nr) of each item is defined by the following equation (1):

$$nr = \frac{1}{n}\sum_{1}^{n} w \,, \qquad (1)$$

In this equation, n is the number of users who evaluated each item, and w is the weight assigned them to each item. The similarity perspective of the evaluator in relation to the models being compared is determined by taking into consideration what he/she finds in the textual or structural characteristics (or by a combination of both).

3.1 Individual Evaluation

At this stage, each evaluator or judge runs each query Q and the system shows up a list of results. Evaluators then express their judgment of similarity of each result against the query. To express such judgment, judges must consider the complete representation of the two business processes (query and result) and their experience in the subject

3.2 Searching for Consensus on Discordant Evaluations

At this stage, each evaluator reviews one by one the relevance judgments issued in the previous stage, and compares them with the judgments that other judges have stated. Thus, evaluators may confront how concordant or discordant their given judgment is against each item, according to the judgment of other evaluators. If evaluators believe that their judgment regarding the set of evaluators is too discordant, they can change their judgment guided by the collective response of other evaluators. For instance, if an evaluator qualified an item as not very relevant in stage 1, but the rest of evaluators (panel of judges) rated it as very relevant, that assessment can make the evaluator reflect on his/her judgment and change his/her decision. This feedback allows judges to have an overview of the evaluation made of each item by all the evaluators.

3.3 Results Refinement

At this stage, and after the judges have (or not) changed their positions (taking into account the contribution of the other judges), the results of each query are listed, taking into account a pair of thresholds. Results are thus filtered by values of nr ranging from 50% to 60% (these parameters can be adjusted depending on the desired confidence level), which means that so far they are not considered as truly relevant nor irrelevant and there still exists a high disagreement level among the judges. As in the previous step, judges may re-analyze the pair of BPs and alter their assessment based on the evaluations of the other judges

3.4 Methodology Objectives

A fundamental task for building a BP test repository is the definition of an intuitive evaluation process where the evaluators (judges) collaboratively agree to clarify similarity criteria in the results retrieved by a BP search system. It may thereby determine the quality of these BPs through a consensus view, given that it is almost impossible to access a real BP repository from an organization.

3.5 Measures for the Evaluation of Relevance

Measures for assessing relevance calculate the relevance of the retrieved results of a BP similarity tool in decreasing, gradual, and continuous forms. They measure the gain of a result item based on the position of this item in the ranking, recognizing that the most relevant BPs are most useful if they appear in the top positions of the ranking (Ulrich and Birgitta, 2010).

Graded relevance measures (*Pg* and *Rg,* described below) must be applied in the above to provide a classification (*Ti*) of the BPs returned in the repository, those that are considered similar to a query BP (*Q*) according to different levels of relevance. *Pg* and *Rg* (Tsetsos et al., 2006) take into account the sum of degrees of relevance Among the BPs.

In addition, to measure the quality of the ranking of the results generated by the BP searching mechanism applied on the current evaluation, ANDCG (Average Normalized Discounted Cumulated Gain) and GenAveP' (Generalized Average Precision) (Ulrich and Birgitta, 2010) measures were used as presented and improved in the works of Küster and König-Ries (2008). These measures quantify the quality of the ranking produced by Web services´ retrieval tools, but are fully applicable to the BP searching field.

4 DEVELOPED TOOL

The main purpose of the platform is to provide an infrastructure to integrate a group of judges (evaluators) in a collaborative environment to issue relevance judgments regarding the set of results reported for different queries by a BP searching engine. The platform enables the implementation of any BP search engine that integrates the required features to capture data in the indexing and searching interface. All the functionality is provided through a Web user interface. In this sense, the platform allows manual and intuitive comparison of the BPs within a given repository, according to each query. Next we describe the architectural components of the tool.

An architecture composed by three layers was defined for the development of the application (see Figure 1). This architecture provides the following advantages: flexibility, scalability and facilitates the construction and maintenance of the platform. These layers are described below.

Presentation Layer: This layer includes a simple and usable user-centric Web interface that can be accessed using any Web browser. Therefore, this interface provides a visual functionality for evaluators (judges) to execute each query, and additionally specifies the relevance level through a consensus view in a collaborative environment for each one of the searching results classified and sorted sequentially in a list.

Business Logic Layer: this layer comprises business rules and processes related to the functionality offered by the system and that are implemented at this layer. For instance: executing each evaluation phase, running query options in the search engine (which may be a list of the *M* BPs from the repository or a short list of *T* <=*M* BPs that relies on a searching tool to reduce the judges efforts), evaluating retrieved items, giving relevant judgment, calculating relevance, providing a chat service for users, among others.

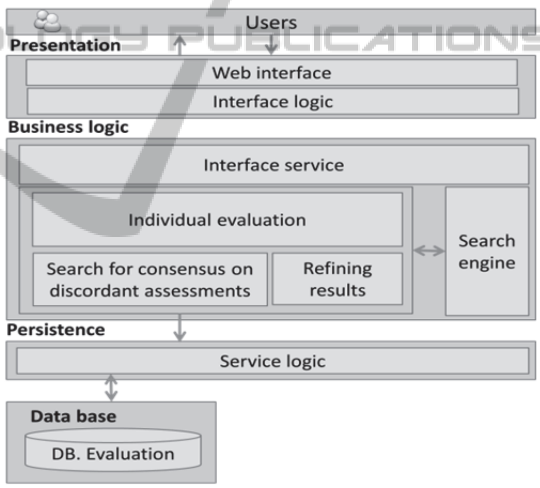

Figure 1: Web Application Architecture.

Persistence Layer: this layer provides the functionality for flexible storing: BP models in an XML representation; BP models to be used as queries; evaluation data of the judges; and evaluation judgments about each of the retrieved items according the queries. Besides, this layer provides agile and efficient mechanisms to retrieve, access and manage the existing BP models in the repository and the collected information throughout the evaluation process.

Figure 2 depicts the individual evaluation interface that was developed for the evaluation step. The tool was implemented with Java technology, additionally PostgreSQL was used as RDBMS for storing the information managed in the evaluation

5 REPOSITORY BUILT

This section presents the results obtained in the manual comparisons made by the judges using the developed platform and the concordance and the evolution of consensus judgments using the proposed methodology.

5.1 Repository

The current implementation of the repository includes 100 BPs modeled with BPMN (Business Process Modeling Notation). Those BPs were graphically designed by experts of the Telematics Engineering Group of the University of Cauca (Colombia) based on real processes provided by Telco operators in Colombia and examples found in different Web sites (e.g., the TM Forum)(Figueroa 2011). It was not possible to use a real repository of a Telco operator because operators are reluctant to give access to their repositories due to privacy and security policies. This is available in the following link:https://drive.google.com/file/d/0B1J2e8JSqOR2QlBQcENPdXlMMTA/edit?usp=sharing.

5.2 Judge's Profiles

In order to evaluate the proposed methodology, we have counted with 59 people (judges or evaluators), which belong to the Institute of Informatics and to the Business Management School, both of the Federal University of Rio Grande do Sul (Brazil), and to the University of Cauca (Colombia), distributed according to Table 1.

Table 1: Kind of Judges or evaluators.

	Dr.	MSc.	Professional
Institute of Informatics/UFRGS	-	7	14
Business Management School/UFRGS	-	-	33
University of Cauca	2	3	-

5.3 Evaluation Phase

For this phase, a set of 6 BP were defined as query elements, and, for each query, the searching mechanism returned a list of 20 results sorted by the similarity defined within the searching model.

Thus, each judge manually compared the similarity between the query models with each item in the results list, and maked a relevance judgment from the ones established in the methodology (i.e., the Likert scale described in Section 3).

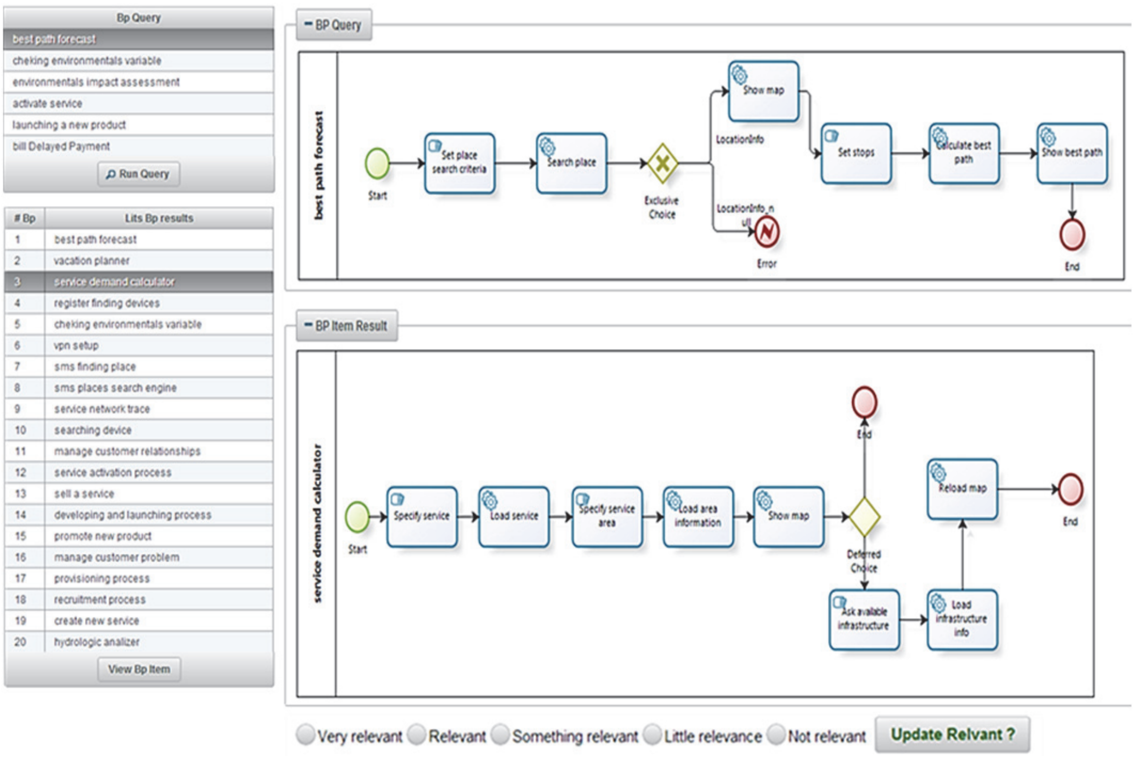

Figure 2: Developed tool, individual evaluation interface.

The evaluation was conducted in this way: each group of judges was gathered to the computers lab at the university they belong to. The evaluation methodology and its aims were explained to the groups once they were met. Subsequently, the operation of the evaluation platform was explained, and the individual evaluation phase was started in a coordinated way. This is because it is necessary to start the searching for consensus on discordant items taking as initial state the whole set of relevance judgments issued by the judges from each group during the evaluation phase

Once the first phase was finished, a period of time was established to complete the other evaluation phases. For this purpose, we have established communication via mail as a reminder element on the completion of the final evaluation stages.

According to the above, each judge provided an average of 360 manual comparisons, in that sense, the total of manual comparisons made by the judges was around 21,240.

5.4 Methodology Application on the Repository

Comparisons made by the judges in a manually way at each one of the stages (St1- Individual, St2- Searching for consensus on discordant evaluations, and St3-Results refinement) based on standard deviation allow an overview of the concordance level between them. In Table 2 we present the concordance values between judges for the items evaluated at each query stage. This value is represented by grouped standard deviation values, which measures the relevance levels dispersion which are classified within the range values previously presented.

In relation to the application of the methodology on the repository, the following average concordance (AVG) values between the judges were obtained: 0.284 for stage 1, 0.256 for stage 2 and 0.250 for stage 3. These values indicate that these relevance judgments are not widely dispersed and therefore do not differ much. When judges progress through the evaluation stage, these values are lower and tend to commonalities showing the force of the proposed methodology.

In addition, it has a 9.7% of concessive improvement in (MCF) between stage 1 and stage 2, and 2.4 % between stage 2 and stage 3 for each query, confirming that stage evaluations allow to better refine the repository (results by each query).

This allows us to perceive that the 59 judges improved their consensus at 11.8%, unlike if they would have done individually. In this sense, the repository gets 11.8% of general concessive improvement (MCG) making it more "ideal" than required at stage 1.

Besides, the collaborative evaluation methodology and the developed tool minimize the re-evaluation work in stages 2 and 3.

Consequently, the collaborative evaluation methodology and this tool improve the repository quality, increasing its usefulness.

In addition, the Pearson correlation coefficient was used to calculate the concordance level between judges in each of the stages (St1 to St3) for each query. For this, we took as population the relevance judgments executed by the evaluators (judges) to each item in the list. The Figure 3 shows that the correlation becomes stronger as the stages advance and evaluation goes forward. Consequently, *Q*1 scored the lowest concordance level between stages 1 and 2, achieving 83%. Similarly, between stages 2 and 3, it scored 87%. Moreover, *Q*6 scored the highest concordance degree between stages 1 and 2,

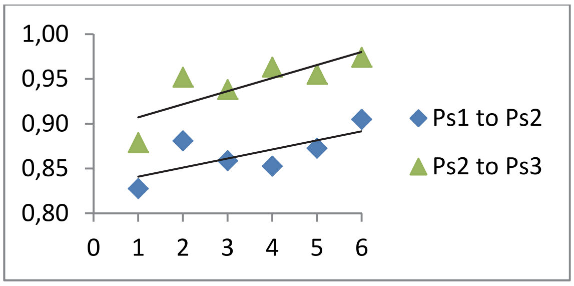

Figure 3: Concordance between evaluators to each stage.

Table 2: Standard deviation value by each relevance judgment per phase.

Measure	Q1			Q2			Q3			Q4			Q5			Q6		
	St 1	St 2	St 3	St 1	St 2	St 3	St 1	St 2	St 3	St 1	St 2	St 3	St 1	St 2	St 3	St 1	St 2	St 3
AVG	0,31	0,27	0,26	0,29	0,27	0,26	0,28	0,25	0,25	0,27	0,24	0,23	0,28	0,26	0,25	0,28	0,25	0,25
MCF		11,2%	2,8%		8,6%	3,7%		9,4%	2,1%		10,3%	1,5%		7,4%	2,3%		11,1%	1,9%
MCG			13,6%			11,9%			11,3%			11,6%			9,6%			12,8%

achieving 90%. In the same way, between stages 2 and 3, it scored 97%, showing that concordance level between judges is a growing correlation (very high and positive).

6 CASE STUDY

This section presents the outcomes of applying the methodology on the repository built using a BP searching mechanism. In our case, we have used a BP model searching mechanism that uses linguistic information (activity name, activity type and description) and structural information; it is called a MultiModalSearBP model that is described as follows.

6.1 BP Searching Model Applied

The discovering process applies a searching strategy that integrates linguistic and structural information contained in the BPs, thus allowing us to increase the effectiveness and relevance of the searching results. The MultimodalSearchBP architecture consists of three layers, described below.

Parsing Layer: This layer has a parser that transforms BPs from its original format XPDL (XML Process Definition Language) to a vector representation, where each BP is considered a term's matrix consisting of a linguistic component and other structural.

Indexing Layer: This layer gives a weight to the linguistic and structural components in order to create a multimodal search index consisting of the linguistic matrix component (*MC*) and the matrix structural component (*MCd*) as follows: $MI = \{MCd \cup MC\}$, and the index stores the physical file location of each of the models stored in the repository.

Query Layer: This layer is responsible for allowing BP's search from three querying options: linguistic, structural, and multimodal query (Ordoñez 2013).

6.2 Analysis of the Results

In this section, the results obtained using the search engine on the built repository are presented.

For this, it is necessary to create an outcome list with the items considered as relevant by the judges for each query, which is sorted from highest to lowest depending on the relevance level (*nr*), achieved in manual evaluation.

Then, the resulting list generated by this BP searching mechanism is compared to the resulting list considered as relevant by the judges on that query. In Figure 4Figure 4, the evaluated searching model achieves a grated precision (*Pg*) average that ranges from 57% (minimum) to 85.2% (maximum). This model combines structural and linguistic criteria present in the BPs, over text processing algorithms capable of reducing the probability of retrieving irrelevant results (false positives).

Regarding to graded Recall (*Rg*), it ranges between 34% and 56%. This is because the number of results returned by each query is limited to twenty BPs. This limitation is inspired in the Web search domain, where users only are focused on the first ten or twenty results in the answers set. Therefore, this indicates that the model can get false negatives (lose relevant business processes in the ranking), but at the same time increases accuracy by reducing the number of false positives.

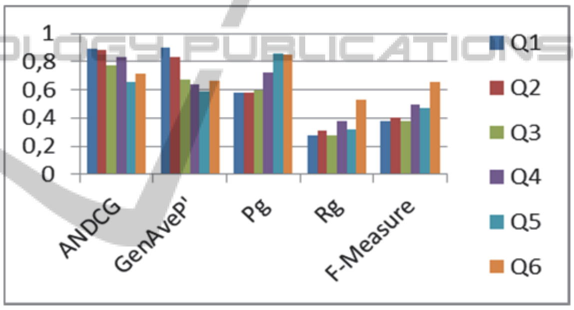

Figure 4: Evaluation measures.

About to the effectiveness of the searching model, it is characterized by the performance obtained in the rankings. In that sense, F-Measure allows observing the harmony of Pg and Rg results, and, in the searching model applied, it obtained average values between 36% and 47%. Regarding to the results ranking, ANDCG demonstrates that the ranking generated by the model used has high quality, because it places a representative number of relevant elements at the beginning of the ranking, reaching an average range between 79% and 88%. As explained before, the difference between GenAveP and ANDCG' measures is that the last one possesses a factor that evaluates the elements retrieved to the bottom of the ranking with a higher value. In these cases, the model reached an average value between 71% and 88. The graded measures provide a more intuitive and flexible evaluation. They also reduce the influence of inconsistent judgments among evaluators

7 CONCLUSIONS AND FUTURE WORK

In this paper, we have established a methodology for the collaborative construction and evaluation of BP repositories. For this purpose, we used a BP searching mechanism applying graded measures to determine the relevance degree of the retrieved elements. Consequently, this allowed the demonstration of the usefulness of the responses and their relationship to queries submitted by users. These responses serve as the most appropriate responses for evaluating and comparing searching mechanisms that use the same repository.

The collaborative evaluation allows judges to have an overview of the relevance judgments issued by each judge on elements retrieved in the results list. As a result, judges can compare the concordance or discordance in the relevance judgment issued for an evaluated item and thus corroborate or change their assessment.

The data shows that there are some differences in the points of view of the evaluators. While most experts considered the items ordered at the top of the result list (1, 2, 3, 4) as relevant or very relevant, a minority (10%) of these were considered as not relevant or irrelevant. This is because the latter took into account only one part of the evaluation process (linguistic or structural), or simply because the comparison between the BP query and each one of these results was performed superficially, which may have been due to fatigue as a result of the huge number of evaluations performed.

The application methodology proposed serves as the basis for the generation of stable evaluations of BP repositories, which are thus more maintainable and reusable. In addition, as a secondary contribution, the BP repository that was used in our evaluation can be seen as an open access repository that will be shared, expanded with new models BP, and can be used in future researches by any actor interested in the area of BP management.

As a future work, it is aimed to expand the evaluation methodology by manually creating groups or families of BPs with those BPs considered as truly relevant in each one of the queries. This allows group representation of thematic topics or structural patterns of the BPs within the repository.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the collaboration of the judges belonging to the Institute of Informatics and of the Business Management School of the Federal University of Rio Grande do Sul (Brazil) and the Department of Telematics of the University of Cauca (Colombia). They also acknowledge the Institute of Informatics for sharing their infrastructure. Finally, we would like to state that this research was partially supported by CAPES and CNPq, Brazil.

REFERENCES

Becker, M. and R. Laue (2012). "A comparative survey of business process similarity measures." Computers in Industry 63(2): 148-167.

Chris J. Turner, A. T., Jorn Mehnen (2010). "A Genetic Programming Approach to Business Process Mining."

Dijkman, R., M. Dumas, et al. (2011). "Similarity of business process models: Metrics and evaluation." Information Systems 36(2): 498-516.

Dijkman, R., M. Dumas, et al. (2011). "Similarity of business process models: Metrics and evaluation." Information Systems 36: 498-516.

Koschmider, A., T. Hornung, et al. (2011). "Re-commendation-based editor for business process modeling" *Data & Knowledge Engineering* 70: 483-503.

Kunze, C. R. a. M. (2013). "An Extensible Platform for Process Model Search and Evaluation." Business Process Management Demos 2013: Beijing, China.

Kunze, M. and M. Weske (2012). "An Open Process Model Library." Business Process Management Workshops, BPM 2011 International Workshops Clermont-Ferrand, France, August 29, 2011 Revised Selected Papers, Part II: 26-38.

Küster, U. and B. König-Ries (2009). "Relevance Judgments for Web Services Retrieval - A Methodology and Test Collection for SWS Discovery Evaluation." 2009 Seventh IEEE European Conference on Web Services: 17-26.

La Rosa, M., H. a. Reijers, et al. (2011). "APROMORE: An advanced process model repository." Expert Systems with Applications 38: 7029-7040.

Ordoñez, C. F., Juan Carlos Corrales, Carlos Cobos (2013). "Multimodal Model for Business Process Search." Submitted to : Information-Sciences.

Rosa, L., H. M. Arthur, et al. (2010). "QUT Digital Repository: http://eprints.qut.edu.au/ This is the author version published as:."

Škrinjar, R. and P. Trkman (2012). "Increasing process orientation with business process management: Critical practices'" International Journal of Infor-mation Management.

Tsetsos, V., C. Anagnostopoulos, et al. (2006). "On the Evaluation of Semantic Web Service Matchmaking Systems." 2006 European Conference on Web Services (ECOWS'06): 255-264.

Yan, Z., R. Dijkman, et al. (2012). "Business process model repositories – Framework and survey." Information and Software Technology 54: 380-395.