7-1-2013

# Identifying Output Interactions Among Is Projects - A Text Mining Approach

Christian Meier
*University of Paderborn, Paderborn, NRW, Germany*, christian.meier@wiwi.uni-paderborn.de

# IDENTIFYING OUTPUT INTERACTIONS AMONG IS PROJECTS – A TEXT MINING APPROACH

Meier, Christian, University of Paderborn, Warburger Str. 100, 33100 Paderborn, Germany, Christian.Meier@wiwi.uni-paderborn.de

## Abstract

*The information systems (IS) literature provides anecdotal as well as empirical evidence for the presence of output interactions amongst IS projects, and their business impact. A number of sophisticated optimization models have been suggested for the consideration of output interactions when selecting IS project portfolios, but usually, the necessary data required for their application in business practice is not available at the planning stage. The literature currently does not offer techniques on how to identify output interactions at the planning stage - a gap which we attribute to the semantical nature of output interactions. We contribute to filling this gap by applying semantic clustering – a technique originating in the text mining literature – to the field of information systems project portfolio selection. A prototypical decision support system is developed that uses latent semantic analysis and hierarchical clustering to identify potential output interactions among information systems project proposals based on semantic similarities within their goal descriptions. This research-in-progress paper focuses on the design of the prototype developed and argues that latent semantic analysis presents a very promising technique for the identification of output interactions among information systems projects.*

*Keywords: Information Systems, Project Portfolio Selection, Project Interactions, Latent Semantic Analysis, Semantic Clustering.*

# 1 Introduction

The selection of the right information systems (IS) projects to form an adequate project portfolio has become an increasingly "important and recurring activity in many organizations" (Archer and Ghasemzadeh, 1999). An often neglected requirement in this selection process is the consideration of project interactions. Three types of interactions can be distinguished: (1) overlap in project resource utilization, (2) technical interdependencies, and (3) effect interdependencies (referred to as output interactions[1] in the following) (introduced by Aaker and Tyebjee, 1978, adopted by, e.g., Santhanam and Kyparisis, 1995; Lee and Kim, 2001; Eilat et al., 2006). Making the effort of considering these types of interactions may constitute "valuable cost savings and greater benefits" to an organization (Santhanam and Kyparisis, 1996). The requirement to identify and account for interactions among IS projects in order to avoid making unfavorable project portfolio selection (PPS) decisions may be challenging and time consuming, yet it is also very important (Lee and Kim, 2001). There is anecdotal as well as empirical evidence for the existence of output interactions. For example, based on a data set of 623 U.S. firms, Aral et al. (2006) name complementarities between the implementation of Enterprise Resource Planning, Customer Relationship Management, and Supply Chain Management Systems as an explanation of performance gains. On a data set of 927 German firms, Engelstätter (2009) finds similar results. He observes positive effects among three enterprise software systems when they are used together and attributes this observation to possible complementary effects occurring between them. Hence, in the following, these effects are referred to as *complementary output interactions*. Besides complementary output interactions, ESI International (2009) reports from a global survey among 470 project and program management professionals that "71% of respondents report redundancies and conflicts in respect to project priorities". In the following we refer to these redundancies in the project portfolios as *competitive output interactions*. Both complementary and competitive output interactions may explain why the business value impacts of projects are non-additive (see, e.g., Fox et al., 1984; Eilat et al., 2006). While the aforementioned studies investigate the existence and impact of output interactions from an *ex post* point of view, to the best of our knowledge no research has been conducted so far that aims at the *ex ante* identification of output interactions. Considering the reported effects and their expected business value impact, an ex ante consideration of output interactions could substantially affect the decision on a portfolio selection.

Numerous articles can be found in the literature that already incorporate output interactions into Operations Research (OR) decision models (e.g., Aaker and Tyebjee, 1978; Santhanam and Kyparisis, 1996; Lee and Kim, 2001). However, the time-consuming identification of output interactions is mostly left unsupported with the portfolio planner. This severely hampers the application of these models in business practice. The lack of contributions to the identification of output interactions can be attributed at least partly to the rather semantic nature of output interactions. In contrast to, e.g., resource requirements, a project's planned outputs and goals tend to be formulated in a textual and less structured form. In addition, the effects of output interactions become visible only after the corresponding projects have been conducted, whereas the effects of overlap in resource utilization or technical interactions may already become apparent during project execution. However, indications for potential connections between project goals may already be contained in the textual descriptions in the project proposals at the portfolio's planning stage. These descriptions, which are often couched in informal language, serve the purpose of communicating project goals to co-workers and decision makers. Thus, we expect output interactions to be found within the semantics of these descriptions. To date these interactions have to be identified manually by domain experts. Especially in large project

[1] In the following, we speak of an output interaction if within the outputs of two or more projects there is an overlap in the provided project goals or services with the result that the business value impact of projects is non-additive.

environments where potentially a large number of output interactions may occur, their manual identification by a human expert can become very challenging and time consuming. In various application domains, latent semantic analysis (LSA) (Deerwester et al., 1990), an information retrieval technique from the text mining literature, could be successfully applied to identify semantic similarities among a set of text documents. Typically, LSA is applied in the context of search engines (e.g., Berry et al., 1995) with the goal to identify documents best matching a certain search query. In this paper we constitute a starting point for a more detailed ex ante identification of output interactions within IS project portfolios by applying LSA to the domain of IS project portfolio selection (IS PPS). Thus, the main contribution of this paper is the development of a prototypical Decision Support System that confers well established concepts from the text mining and information retrieval domain to the field of IS PPS. In a cumulative research tradition, we base our prototype on an approach called semantic clustering presented by Kuhn et al. (2005), which uses LSA for the identification of semantic topics in source code, and adapt it to the new conditions arising from the application domain of IS PPS. Grounded in the Design Science Research paradigm (Hevner et al., 2004), the prototype is an instantiation of the LSA concept and draws upon design knowledge from the field of text mining and information retrieval. We contribute to the literature on PPS by addressing the following research question: *How can the identification of potential output interactions in IS project portfolios be adequately supported by semantic clustering?* This knowledge contribution to the field of IS PPS is located in the "exaptation" quadrant in the framework presented by Gregor and Hevner (2013).

## 2 Background

Our research is based on two different strands of literature: The literature on interactions in PPS, and the literature on text mining techniques for the identification of semantically similar topics in text documents. The former provides the theoretical foundations concerning the importance of project interactions when selecting appropriate project portfolios (e.g., Santhanam, R., Kyparisis, 1996; Lee, J.W., Kim, 2001; Eilat et al., 2006) and defines different interaction types (Aaker and Tyebjee, 1978; Kundisch and Meier, 2011a). Further, it offers valuable insights into the design of sophisticated optimization models incorporating the different types of interactions (e.g., Santhanam, R., Kyparisis, 1996; Lee and Kim, 2001).While all of these approaches provide very useful techniques for modeling and solving PPS problems under consideration of interactions, they have been built under the (implicit) assumption that the necessary information for identifying and assessing interactions is available to the planner. For output interactions especially, this assumption is rarely met in practice. As discussed above, planned outputs and goals tend to be formulated in a textual and rather unstructured form. Problems of polysemy and synonymy within the textual descriptions additionally hamper the IS-supported ex ante identification of output interactions. Our approach has its methodical roots in the text mining literature, which provides techniques that may help to overcome some of the problems mentioned above. This strand of the literature focuses on how to extract information from textual data automatically. So-called text classification approaches (see e.g., Manning et al., 2009) are of particularly high relevance to our research and can be divided into *supervised* and *unsupervised* learning approaches. Despite their (generally) better retrieval results, the need of human-supervised training severely restricts the applicability of supervised approaches for identifying output interactions automatically. Some highly sophisticated software-tools exist (e.g., Leximancer (Smith and Humphrey, 2006), Rubryx: Software and documentation: http://www.sowsoft.com/rubryx/), which try to alleviate these problems. However, to the best of our knowledge, all of these tools require a learning phase with a sample data set and a priori knowledge of the categories relevant to the document classification. We expect such knowledge typically not to be available to the planner of an IS PPS. Thus, an unsupervised categorization approach seems to be better suited to solve the problem at hand. The articles that are the most closely related to our work apply LSA (Deerwester et al., 1990) for mapping readers to documents based on their background knowledge of the documents' topics (e.g., Wolfe et al., 1998), for the identification of related topics in software source code documents (e.g., Maletic and Valluri, 1999; Kuhn et al., 2007), and for the determination of "helpfulness" votings in

online user reviews (Cao et al., 2011). Because of similarities in their problem structure, for the purposes of our research the article by Kuhn et al. (2007) is of particular interest. The authors propose an approach called semantic clustering to identify similarities among variable identifiers in software source code. They employ LSA and clustering to group together source code documents with similar vocabulary. Kuhn et al. apply their technique to two different case studies with mixed results. The comparably small size of the processed documents as well as the size and quality of the vocabulary in the source code documents lead to difficulties in the application of semantic clustering to their application domain. The authors state that better results are achieved with larger documents, the use of natural language instead of artificial identifier names as well as a larger vocabulary. In IS PPS these conditions are widely met, which constitutes IS PPS as a promising field of application for semantic clustering. Therefore, in a cumulative research tradition, we adapt semantic clustering presented by Kuhn et al. (2007) and apply it to the domain of IS PPS.

# 3 Prototype Design

A first assessment of the application domain of IS PPS suggests that the basic conditions for a successful application of semantic clustering to IS PPS appear to be met. Project proposal documents typically serve the purpose of communicating projects goals and requirements within an organization and are usually formulated in natural language. A first manual investigation of a small set of project proposals from the IS domain has highlighted that project proposals potentially contain valuable information about output interactions. However, the most interesting information is often embedded within the semantics of the proposals. The same project goals may be expressed in many different ways by different individuals so that a simple comparison of the words used to describe these goals often will not be sufficient for an automated identification of output interactions. Thus, important information with regard to output interactions may not be identified by simply comparing key words in different proposals. In other contexts, LSA has demonstrated its ability to overcome these difficulties and to identify the semantic topics in a set of documents (Landauer et al., 1998). This is achieved by breaking the large vocabulary from the candidate documents down into a considerably smaller set of factors which can be interpreted as linguistic topics. Based on these factors, the proposal documents are clustered and adequately presented to the planner. We expect output interactions to be found particularly among the documents that are clustered together. While this research-in-progress paper is mainly concerned with the design of the prototype, the validation of this hypothesis will be the subject of a full research paper. In this paper we focus on the design of our procedural approach which can be divided into five conceptual phases (see Fig. 1). The five phases and the necessary adaptions to the approach described by Kuhn et al. (2007) are briefly discussed in the following. We extract the goal description from each project proposal document as input for our analysis and parse it into a list of words. The vocabulary in the documents originates in natural language, which favors the application of semantic clustering.

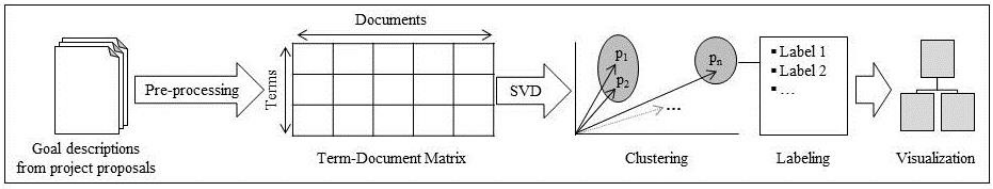


*Figure 1. Identification process.*

Even after this process the project proposal documents still retain a considerable amount of noise due to different linguistic styles of the applicants, words with low semantic relevance, the frequent use of domain specific terms and potentially varying document lengths. As the output quality strongly relates to the quality of the inputs (Kuhn et al. 2007), we implemented an elaborate *pre-processing* to improve input quality. We remove numbers, special characters and single letters and subject the proposal documents to a stemming process (e.g., ‘systems’ is reduced to ‘system’) using the ‘NHunspell

Framework' and the free 'Open Office dictionary'. To remove words which occur frequently but have low semantic relevance, we implement a comprehensive stop word list as well. This list already contains approx. 1.000 generic words (e.g. 'the', 'and', 'of'). In an organizational context, it can be expected that within the proposal documents domain- or company-specific words, abbreviations and phrases (e.g., company or department names, acronyms for company initiatives) have been assimilated into the corporate language to a certain degree. These words and phrases may not contribute to the identification of semantic similarities and have to be identified and added to the stop words list by the portfolio planner in order to improve input quality. The resulting set of words is then arranged into a *term-document matrix*, where the rows represent the terms and the columns the documents. The cell entries represent the raw occurrence of a specific term in a given document. Finally, assuming the goal descriptions differ in length, potential contortions are handled by a widely used normalization and weighting procedure (Dumais, 1991). After pre-processing, *SVD* is applied to the processed term-document matrix to reduce the noise in the data by reducing the number of factors which provide the basis for the document clustering later on. The result of the application of SVD is an approximation of the original term-document matrix that is reduced by noise in the input data and thus, it can be interpreted as a "better model of the text corpus" (Kuhn et al., 2007). The reduced matrix can be represented by a vector space model in which the similarity between two documents can now be established by calculating the angle (usually the cosine) between their corresponding vectors.

*Clustering* represents the key feature of our prototype. Typically, in IS PPS the number of output interactions (and thus the number of clusters) is not known ex ante to the portfolio planner. Thus, we are handling a so-called *unsupervised categorization* problem. Popular clustering algorithms as, e.g., k-means clustering, are not applicable. We therefore implemented an agglomerative hierarchical clustering, which generates a tree-shaped dendrogram. It produces a "hierarchical representation in which the clusters at each hierarchy-level are created by merging clusters at the next lower level" (Hastie et al., 2011). In each step, the clusters exhibiting the highest semantic similarity are merged. This form of visualization enables the portfolio planner to facilitate a better understanding of the relationship structure between the project proposals by presenting the underlying hierarchy of the clustering decisions, instead of being confronted with a single, non-transparent solution. In large project environments, the tree structure may become incomprehensible. It may be helpful for the planner to get an idea which hierarchy level represents a good clustering solution and to only present her with a relevant excerpt of the tree structure. Even if the optimal number of clusters is unknown, numerous techniques can be found in the clustering literature that can be helpful for this task. Milligan and Cooper (1985) provide an overview of 30 heuristic stop criteria to calculate a good clustering based on the coherency *within* and the separation *between* clusters. Therefore, we implemented the Calinsky and Harabasz (1974) index, which performed best in this study, into our approach. While this often may not result in the best possible clustering level for the identification of output interactions from an ex post point of view, we are at least able to suggest a satisfactory hierarchy level based on which the planner can start further analysis. In future research, the visualization of the results as well as a comparison of the performance of different stop criterions with respect to the field of IS PPS have to be thoroughly evaluated. Within clustering, the proposal documents have been grouped based on the semantic topics they share. These topics represent rather abstract linguistic concepts derived from an aggregation of the actual vocabulary used in the documents. To be helpful for the planner, we now have to identify the actual vocabulary from our proposal documents which best defines the topic for the corresponding cluster. Therefore, based on the weighting formula presented in Kuhn et al. (2007) each cluster in the clustering hierarchy is labeled with the *n* most relevant terms from the vocabulary which best describe the topic of that cluster. In a small pre-test we have observed that the number of these top words that are necessary to understand the underlying semantic topic varies from cluster to cluster. Therefore, in addition to the weighting formula of Kuhn et al. (2007), we have already implemented two proprietary labeling strategies as well as a parameterized input for the number of top words the clusters are labeled with. The evaluation of how many top words are adequate in our application domain and which of the labeling strategies provides the best results will form the subject of our future work.

# 4 Summary, Discussion and Future Research

In the literature, there is anecdotal (e.g., Aaker and Tyebjee, 1978) as well as empirical (e.g., Aral et al., 2006) evidence for the existence and the business impact of output interactions among IS projects. While a number of sophisticated optimization models have been suggested which already provide for the consideration of output interactions when selecting IS project portfolios, the necessary data required for their application in business practice is usually not available to the planner. We identify a lack of techniques in the literature on how to identify output interactions ex ante to the portfolio selection process, and attribute this gap partly to the semantical nature of output interactions. With this paper, we contribute to filling the identified gap by importing semantic clustering – a technique originating in the text mining literature – into the field of IS PPS. We develop a prototypical DSS that uses LSA and hierarchical clustering to identify potential output interactions among IS project proposals based on semantic similarities within their goal descriptions. This paper discusses the design of the developed prototype and argues that LSA offers a very promising technique for the identification of output interactions among IS projects. For practitioners, the resulting prototype may serve as a tool to identify output interactions in a structured and potentially more rigorous way and to include them into their portfolio decisions. We expect our approach to perform particularly well for the identification of competitive output interactions, as this type of interaction seems to be less subtle than complementary output interactions. In addition, the hierarchical representation chosen in this paper may highlight relationships within an organization's project landscape which may not have been recognized explicitly before. For researchers, the presented approach may constitute a starting point to incorporate the identification of output interactions into new or existing approaches. To develop this work into a full research paper, several points have to be addressed in future work. As required in design science research, the applicability of the approach has to be evaluated thoroughly. In line with the framework presented by Pries-Heje et al. (2008), we plan to conduct an *ex post* evaluation in a real world setting comprising two steps. In a first step the information retrieval quality of the presented prototype will be evaluated against other state of the art approaches, using the pre-classified *Reuters-21578* standard test set for categorization procedures. To obtain the highest possible comparability, we use the evaluation framework suggested by Massey (2005) and the F1 quality measure, which is state of the art for text classification approaches (see Manning et al. 2009). Second, we plan to apply the approach to a large real world data set of IS project proposals. Therefore, a large data set will be acquired (approx. between 50 and 150 project proposals) as well as a reference solution from domain experts. The solution provided by our approach will then be thoroughly evaluated against the experts' reference solution. Further, the development of the prototypical DSS discussed above comes along with several design choices. These choices have to be assessed against a number of alternatives in the future in order to evaluate the applicability of the technique presented to the problem at hand. It has to be determined how the exclusion of domain- and company-specific stop words and phrases influences the solution quality of the approach and how these stop words may be identified automatically by the prototype. In addition, so-called relevance feedback (Dumais, 1991) may be implemented which allows the planner to define which of the identified interactions are relevant and which can be neglected in a further iteration. The labels of the irrelevant clustering results could be added to the stop word list and be ignored in further iterations. Finally, different stop criteria for the clustering procedure as well as the labeling quality have to be evaluated in cooperation with domain experts.

# References

Archer, N.P., Ghasemzadeh, F.(1999). An integrated framework for project portfolio selection, In: International Journal of Project Management, 17(4), 207-216.

Aaker, D.A., and Tyebjee, T.T. (1978). A Model for the Selection of Interdependent R&D Projects. In IEEE Transactions on Engineering Management, 25(2), 30-36.

Aral, S., Brynjolfsson, E., and Wu, D.J. (2006). Which came first, IT or Productivity? The Virtuous Cycle of Investment and Use in Enterprise Systems. In Proceedings of the 27th International Conference on Information Systems, AIS, Milwaukee.

Berry, M.W., Dumais, S.T., and O'Brien, G.W. (1995). Using linear algebra for intelligent information retrieval, SIAM Review, 37(4), 573–597.

Calinsky, R.B., and Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3, 1-27.

Cao, Q., Duan, W., Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems, 50, pp.511-521.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391-407.

Dumais, S.T. (1991). Improving the retrieval of information from external sources. In Behavior Research Methods, Instruments and Computers, 23, 229–236.

Eilat, H., Golany, B., Shtub, A. (2006). Constructing and Evaluating Balanced Portfolios of R&D Projects with Interactions: A DEA based Methodology. In European Journal of Operational Research, 172, 1018-1039.

Engelstätter, B. (2009). Enterprise Systems and Labor Productivity: Disentangling Combination Effects, In International Journal of Engineering Research and Applications, (forthcoming).

ESI International Inc. (2009). View from the Ground: The Project Manager Perspective on Project Portfolio Management Effectiveness. http://www.esi-intl.co.uk/resource_centre/white_papers.

Fox, G.E., Baker, N.R., Bryant, J.L. (1984). Economic Models for R and D Project Selection in the Presence of Project Interactions. In Management Science, 30(7), 890-902.

Friedman, J.H., Hastie, T., Tibshirani, R., (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition, Springer-Verlag, Heidelberg.

Gregor, S., Hevner, A.R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. MIS Quarterly, 37(2), 337-355.

Hevner, A.R., March, S.T., Park, J., and Ram, S. (2004). Design Science in Information Systems Research. In MIS Quarterly, 28(1), 75-105.

Kuhn, A., Ducasse, S., and Gîrba, T. (2007). Semantic clustering: Identifying topics in source code. Information and Software Technology, 49, 230–243.

Kundisch, D., and Meier, C. (2011a). IT/IS Project Portfolio Selection in the Presence of Project Interactions - Review and Synthesis of the Literature. In Wirtschaftsinformatik Proceedings, 477-486.

Landauer, T.K., Foltz, P., Laham, D. (1998). An introduction to latent semantic analysis, Discourse Processes, 25(2-3), 259-284.

Lee, J.W., Kim, S.H. (2001). An Integrated Approach for Interdependent Information System Project Selection. In International Journal of Project Management, 19, 111-118.

Maletic, J.I., Valluri, N. (1999). Automatic software clustering via Latent Semantic Analysis. In Proceedings of the 14th IEEE International Conference on Automated Software Engineering, FL.

Manning, C.D., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge University Press, Online edition.Massey, L. (2005). Evaluating and comparing text clustering results. In Proceedings of International Conference on Computational Intelligence.

Milligan, G.W., and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.

Pries-Heje, J., Baskerville, R., Venable, J. (2008). Strategies for Design Science Research Evaluation. In European Conference on Information Systems Proceedings, Paper 87.

Santhanam, R., Kyparisis, G.J. (1996). A Decision Model for Interdependent Information System Project Selection. In European Journal of Operational Research, 89, 380-399.

Smith, A.E., Humphreys, M.S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. Behavior Research Methods, 38(2), pp. 262-279.

Wolfe, M.B.W, Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W., Landauer, T.K. (2009). Learning from text: Matching readers and texts by latent semantic analysis. Discourse Processes, 25(2-3).