Intra-Organizational Boundary Spanning: A Machine-Learning Approach

Emergent Research Forum

Wietske van Osch
Michigan State University
Vanosch@msu.edu

Charles Steinfield
Michigan State University
Steinfie@msu.edu

Yanjie Zhao
Michigan State University
Zhaoyanj@msu.edu

Introduction

In recent years, the ability to mine, manage, and examine big data has sparked a strong interest among scholars and managers to leverage data science and machine-learning approaches for enhancing the efficiency and effectiveness of various areas of knowledge management. The success of today's enterprises increasingly depends on the efficiency and quality of their cross-boundary knowledge flows and processes (Marrone, 2010). Various information systems, specifically emerging enterprise social media (ESM) technologies, are used to increase the transparency and openness of knowledge flows with the aim of enhancing team effectiveness, collaboration, knowledge sharing, and innovation.

Today, the proliferation of ESM technologies in organizational contexts has profound implications for such boundary-spanning knowledge flows in organizational teams. Social media encompass a range of information and communication tools (ICTs) for supporting interaction, collaboration, and co-creation, such as blogs, content communities, and social network sites (Leonardi, Huysman, and Steinfield, 2013; Treem and Leonardi, 2012). Studies of organizational social media use suggest that these systems have the potential to enhance boundary-spanning knowledge flows by enabling the identification of and interaction with relevant external individuals and information (cf., DiMicco et al. 2008; 2009; Steinfield et al., 2009).

In this paper, we develop and test a machine-learning algorithm for detecting three distinct types of boundary spanning drawn from a series of earlier studies on project teams, using content data from an ESM platform of a large multinational corporation. The three boundary-spanning activities include representation, coordination, and general information search (Ancona and Caldwell, 1992), all of which have been associated with distinct performance benefits, both for the teams performing these activities and the organization at large. Hence, insights from the proposed algorithm can assist knowledge managers in evaluating and enhancing the likelihood of cross-boundary knowledge flows.

The remainder of this paper is organized as follows. We first review the boundary spanning literature in knowledge management and offer detailed descriptions of the three afore-mentioned boundary-spanning activities and their associated performance benefits. Subsequently, we describe the case organization, data collection, manual data analysis, and the construction of the machine-learning algorithm. We then present the preliminary findings generated by the algorithm and its reliability. Finally, we discuss next steps as well as important implications for research and practice.

Boundary Spanning

One of the most pressing challenges for the field of knowledge management is how to overcome and connect knowledge silos in order to facilitate efficient and effective cross-boundary knowledge flows within the organization (Nonaka and Takeuchi, 1995). This challenge constitutes one important area of knowledge management frequently referred to as boundary spanning.

Boundary spanning–the extent to which communication links units to external sources of information (Tushman and Scanlon, 1981)–is closely related to other popular concepts from social network theory, including bridging or weak ties (Granovetter, 1973) and structural holes and information brokerage (Burt, 1992). The common denominator across these concepts is the importance of establishing and managing external linkages as conduits to critical resources, coordination, and the creation of reputational benefits.

Hence, for managers to detect whether and where knowledge silos exist, they need reliable methods for detecting the existence and nature of boundary-spanning activities or the lack thereof. In this paper, we propose a machine-learning approach that can facilitate the automatic and reliable detection of three distinct types of boundary-spanning activities as identified in the boundary spanning literature, namely representation, coordination, and information search, as follows.

Representation involves the lobbying for the team up the hierarchy in order to create favorable impressions amongst senior managers, hence, is a largely vertical form of boundary spanning (Ancona and Caldwell, 1992). This process is crucial for team performance as the creation of a favorable impression among senior management is a prerequisite for obtaining access to key resources (e.g., reputation, legitimization, higher-level commitment) and financial support needed for successful product development (Grabher, 2004). Representation further benefits management as they stay informed of team progress that can support higher-level planning and resource allocation decisions, which in turn, can help the organization meet external client expectations (cf., Bettencourt et al., 2005).

Coordination involves the facilitation of effective decision-making and design implementation through cross-boundary strategizing, planning, and evaluation; hence it is a horizontal form of boundary spanning (Ancona and Caldwell, 1992; Marks et al., 2001). This process is crucial for team performance as it involves the aligning, negotiating, and monitoring of the efforts of individuals—within and outside the team—in order to accomplish project goals (e.g., delivery deadlines). Hence, coordination is crucial for the efficiency, effectiveness, innovativeness, and flexibility of goal delivery (Mohrman et al., 1995).

General information search involves the general scanning of the external team environment to gain access to relevant information, knowledge, and expertise; hence, is a largely horizontal form of boundary spanning (Ancona and Caldwell, 1992). Target actors of information search activities are often loosely coupled with the focal team (Marrone, 2010). This boundary-spanning process is crucial for team performance as it enables them to gain project-specific expertise and an understanding of trends, opportunities, and threats in the external environment (Hargadon, 1998).

Research Design

Data Collection and Study Context

Data was collected from the ESM platform of a large worldwide provider of workplace products, furnishings, and services. The company has approximately 10,000 employees around the world and is headquartered in the U.S. with offices and divisions in nearly 40 countries in North and South America, Europe, Africa, Asia, Oceania, and the Middle East.

In March 2012, the organization launched an ESM tool based on the Jive Platform. Jive[1] is a provider of corporate social technologies that support business communications and collaborations among employees. Following its global launch in March 2012, the adoption and use of the system has grown substantially, with a total user base of over 9,000 users as of 2014.

For the development of the machine-learning algorithm, we collected content data from the blogs and discussion threads of 463 groups, resulting in a total of 2029 discussions and 6500 threads.

Data Analysis

To ensure the reliable development of the machine-learning algorithm, three graduate students were trained to perform manual coding of the content data, assigning the various posts to categories reflecting

[1] http://www.jivesoftware.com

the type of boundary-spanning activity each contained (or lack thereof). Coding was preceded by an elaborate training session to familiarize the coders with the coding manual and the coding scheme.

The coding manual included five coding categories, namely three categories for each of the three boundary-spanning activities—representation, information search, and coordination—as well as two additional categories for classifying posts that appeared unrelated to boundary spanning. For those activities unrelated to boundary-spanning, coders had to decide whether the activity was related or unrelated (e.g., social) to work.

Following the training, the coders were supervised in the independent coding of 14% of the content to compute interrater agreement. An initial interrater agreement of 89.6% with a corresponding .71 Cohen's kappa (i.e., substantial agreement; c.f. Landis and Koch, 1977) provided confirmation of coding scheme validity and coding process reliability. Following the reconciliation of differences, the remainder of the content data for manual coding was divided across the three coders.

Algorithm Development

Within the next stage, the manually coded data was used to create an algorithm for automated text classification. The problem of text data classification belongs to the area of natural language processing, which is one of the most popular applications of machine learning. Compared to machine-learning problems that deal with numerical data, text data mining and classification is more tedious. In this context, there are three characteristics that make the type of data used in this project specifically complex.

First, text data encompasses a large vocabulary. Since each unique word is treated as one feature, the feature space is highly multi-dimensional. Second, message length is not evenly distributed; while some messages have hundreds of words, others have less than ten. Third, the boundary-spanning categories are not mutually exclusive and so there is a multi-labeling problem; e.g., one post could be simultaneously a representational and an information search activity. In what follows, we describe the step-wise algorithm development process including the steps taken to address the aforementioned complexities.

The first step prior to classifier development included data cleaning and preprocessing, during which we (i) removed the html style format, punctuations, numbers, non-English words and stopwords, and (ii) converted all words to lower case.

The second step included feature selection, which is to choose the most representative words and build an overall dictionary. The words chosen have to be biased, in other words, they should be highly related to one or a couple of the boundary-spanning categories. The words should also appear in a higher frequency for a specific category to reflect high reliability. The total number of words extracted from all the text documents is 13,791, which constitutes a relatively large original feature dataset, thus, the method chosen for feature selection also needs to be relatively efficient in terms of computation time. Based on these considerations, the gini-index equations (Aggarwal, 2014) were applied to feature selection, since the computation for this method is much faster than some other feature selection approaches such as information gain. The equations are described as follows:

$$p_i'(w) = \frac{p_i(w)/P_i}{\sum_{j=1}^{k} p_j(w)/P_j} \tag{1}$$

$$G(w) = \sum_{i=1}^{k} p_i'(w)^2 \tag{2}$$

In equation (1), the $p_i(w)$ is the fraction of class i presence for the word w. That is to say, pi(w) is the conditional probability that a document belongs to class i, given the fact that it contains the word w. P_i denotes the global distribution of the documents belonging to class i in all the documents. Equation (1) computes the normalized probability for each word in a certain class and equation (2) is used to calculate the gini-index value of each word. The higher the gini-index value is, the more representative the word is.

The third step involved the computation of word presence frequencies, which are then sorted in descending order. The final dictionary was eventually built based on the two lists with a threshold which gives the highest accuracy after applying a 10-fold cross validation method.

We next chose a support vector machine (SVM) learning algorithm to develop the prediction model. A support vector machine is a supervised learning algorithm. The idea of this algorithm is to construct a (set of) hyperplane(s) in a high dimensional space, which can be used to separate data samples belonging to different classes. The hyperplane chosen should have the largest distance to the nearest training data points from different classes as a larger margin will lead to lower generalization error.

SVM is particularly well-suited for text categorization for a number of reasons (Joachims, 1998). For example, text data has many features with each unique word as a feature, and SVM deals well with high-dimensional data since SVM offers overfitting protection. Also text data, after being processed, will generate a sparse matrix and SVM is well suited for sparsity.

Although the original SVM algorithm was invented in 1963 as a linear model, a kernel trick was later introduced to be applied to nonlinear classifiers (Boser et al., 1992). In this project, we applied a sigmoid kernel function in the nonlinear classification model, which was determined using cross validation by comparing the performance to other kernel functions including linear and RBF functions.

Preliminary Findings

The total number of discussion messages is 2029, of which 304 were labeled manually by the coders. After removing the meaningless messages, 285 messages were retained and used as the final training samples. Features (i.e., words) were selected based on two lists: the gini-index value and frequency. Initially the threshold of frequency 1, 2, 3 combined with a gini-index threshold of 0.65, 0.70, 0.76, 0.87, 0.90 were applied to the algorithm and corresponding accuracies generated by 10-fold cross validation were compared. Our reason for choosing 10-fold cross validation is based on the consideration of the size of the training sample dataset. With 5-fold cross validation, some featured words that belong to only the testing data will likely be lost, while 20-fold cross validation would be challenging given that the number of testing samples would become too small to obtain steady accuracy.

Eventually, the combination of frequency of 1 with gini-index number of 0.87 provided a better training sample set generating higher accuracy (Table 1). When the frequency threshold was set to more than 1, some of the training sample vectors turned out to be 0 for all features. This may be due to the very short length of some messages, which may not contain any common word from the dictionary.

Single Classifier	Total accuracy (%)	67.0									
	Single accuracy (%)	71.4	69.0	67.9	65.5	64.3	62.1	50.0	86.2	64.3	69.0
Assembled classifier	Total accuracy (%)	79.0									
	Single accuracy (%)	86.2	78.4	85.8	69.6	73.5	85.8	80.1	84.6	72.1	73.5

Table 1. Accuracies of 10-fold Cross Validation of Training Samples

After examining the classification results for individual classes generated by the single classifier, we found that the model was highly biased, assigning over 70% of training samples to the "information search"

category in contrast to only 43% of training samples belonging to "information search". To solve this problem we applied two approaches: 1) adding a cost matrix to the classifier, which was computed based on the distribution of each category, in order to "force" the classifier to give more weight to those minor categories such as "representation" and "work related activity"; 2) predicting two labels instead of one by assembling multiple classifiers obtained previously using different thresholds of gini-index values. This modification is reasonable because some text content actually reflects more than one boundary-spanning category. Accuracy was calculated by accepting the classification result as correct if one in two labels matches the true category. The overall accuracy was improved by 12% to a high of 86.2% (Table 1).

Finally, Table 2 presents the distribution and examples of boundary-spanning activities enacted through ESM. Information search is by far the dominant activity conducted through ESM, accounting for approximately 56.5% of all classified activities. Furthermore, representation and coordination account for ~8% and ~20% of the classified activities. Finally, almost 30% of the activities communicated and enacted through ESM are non-instrumental (i.e., not work-related) activities and a remaining ~7% of activities are work-related but do not fall into one of the boundary-spanning categories.

Categories	Representation	Coordination	Information Search	Work Related Activity	Other activity
Distribution (%)	8.1%	19.8%	56.5%	6.9%	29.0%
Example	Here is a great video showing the work of our team in Vodafone's new workplace.	OK, CDC folks, the cancellation of the innovation center meeting threw us a little curve ball, but here's the revised planning.	Has anyone worked on an innovation centre that they could share?	The Steelcase interns had the opportunity to participate in Chicago yesterday. It was great to see the Steelcase show so full and have such an exciting buzz around it.	Have you hopped on a bike lately??? If so how long and where did you ride?

Table 2. Frequencies and Examples of Boundary Spanning in ESM

Discussion

The next phase of this project involves analyzing the relation between various individual- and group-level variables from the log data and the frequency of occurrence of the boundary-spanning activities conducted by organizational groups. Furthermore, an in-depth, manual exploration of the 7% of activities that did not belong to the representation, coordination, and information search categories is required in order to explore whether additional boundary-spanning activities exist in ESM that have not been previously identified in the knowledge management literature.

Not only can these findings help to advance theories of boundary spanning by providing behavioral insights into individual- and group-level antecedents of boundary-spanning activities, they can further inform managers of those antecedents that are most conducive to successful boundary spanning. Understanding and testing these boundary-spanning antecedents helps to improve the effectiveness of intra-organizational collaboration, knowledge sharing, and innovation.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-1422316. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Aggarwal, C.C. 2014. *A Survey of Text Classification Algorithms*, IBM T. J. Watson Research Center, Yorktown Heights, NY.

Ancona, D. G., and Caldwell, D. F. 1992. "Bridging the boundary: External activity and performance in organizational teams," *Administrative science quarterly* (37), pp. 634-665.

Bettencourt, L. A., Brown, S. W., and MacKenzie, S. B. 2005. "Customer-oriented boundary-spanning behaviors: Test of a social exchange model of antecedents," *Journal of retailing* (81:2), pp. 141-157.

Boser, B.E., Guyon, I.M., Vapnik, V. N. 1992. "A training algorithm for optimal margin classifiers", *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. p. 144.

Burt, R. S. 1992. *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

DiMicco, J.M., Geyer, W., Dugan, C., Brownholtz, B. and Millen, D.R. 2009. "People Sensemaking and Relationship Building on an Enterprise Social Networking Site," *(HICSS)*.

DiMicco, J.M., Millen, D.R., Geyer, W., and Dugan, C. 2008. "Research on the Use of Social Software in the Workplace," *Computer Supported Collaborative Work*. San Diego, CA, USA.

Grabher, G. 2004. "Temporary architectures of learning: knowledge governance in project ecologies," *Organization studies* (25:9), pp. 1491-1514.

Granovetter, M. S. 1973. "The strength of weak ties," *American Journal of Sociology* (78:6), pp. 1360-1380.

Hargadon, A. B. 1998. "Firms as knowledge brokers: Lessons in pursing continuous innovation," *California Management Review* (40), pp. 209-227.

Leonardi, P. M., Huysman, M., and Steinfield, C. 2013. "Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations," *Journal of Computer Mediated Communication* (19:1), forthcoming.

Marks, M. A., Mathieu, J. E., and Zaccaro, S. J. 2001. "A temporally based framework and taxonomy of team processes," *Academy of Management Review* (26:3), pp. 356-376.

Marrone, J. 2010. "Team boundary spanning: A multilevel review of past research and proposals for the future," *Journal of Management* (36:4), pp. 911-940.

Mohrman, S. A., Cohen, S. G., and Mohrman, A. M. 1995. *Designing team-based organizations: New forms for knowledge work*. San Francisco: Jossey-Bass.

Nonaka, I., and Takeuchi, H. 1995. *The knowledge creating company: how Japanese companies create the dynamics of innovation*. New York: Oxford University Press. p. 284

Steinfield, C., DiMicco, J.M., Ellison, N.B., and Lampe, C. 2009. "Bowling online: social networking and social capital within the organization," *Proceedings of the fourth international conference on Communities and technologies (C&T '09)*, ACM, New York, NY, USA, pp. 245-254.

Thorsten, J. 1998. *Categorization with Support Vector Machines: Learning with Many Relevant Features*. Universität Dortmund.

Treem, J. W., and Leonardi, P. M. 2012. "Social media use in organizations: Exploring the affordances of visibility, editablity, persistance, and association," *Communication Yearbook* (36).

Tushman, M. L., and Scanlon, T. J. 1981. "Boundary spanning individuals: Their role in information transfer and their antecedents," *Academy of Management Journal* (24:2), pp. 289-305.