Learning Systems Based Automated Proposal Evaluation

Sam Adhikari
Stanford University
sadhikar@stanford.edu

Advisors from Sysoft Corporation: Arthur E. Miller, Scott R. Palma

Author Keywords

ACM Classification Keywords
Information retrieval: Document representation: Content analysis and feature selection

General Terms
RFP

ABSTRACT

Learning systems based automated proposal evaluation can protect the public procurement agencies and funds from fraud, corruption, unusual, and callas proposal evaluations by evaluators. The impact on the economy from such system is significant. Savings can be in hundreds of billions of dollars across the whole world.

Research suggests that the objective evaluation of a proposal received in response to a Request For Proposal (RFP) is heavily influenced by evaluators' expertise, experience, and knowledge domain. Evaluating proposal from proposers is not easy. Among other challenges, evaluators have to balance objective evaluation decisions against project usability objectives, specific criteria set by the procurement agency, and account for the increasingly diverse range of alternative technical approaches available.

Intelligent proposal evaluation support systems are a potential solution to this problem. They can eventually become automated proposal evaluation systems. However, to intelligently assist evaluators in evaluating proposals, computer systems need to possess *computable representations* of human readable proposals in text. Using machine-extracted text-based features and evaluation judgments from 16 evaluators on 52 proposals, we present preliminary results based on established methodology that suggest that such representation may be possible. This is the first step towards realization of a learning systems based automated proposal evaluation system.

INTRODUCTION AND MOTIVATION

Sysoft eRFP, a group decision support system (http://erfp.sysoft.com) helps Government, and corporate procurement agencies evaluate proposals in response to a Request for Proposal (RFP) [1]. The system is used all over US and abroad for small, medium, and large scale complex procurement projects where best value is desired instead of lowest cost. The procurement projects include IT systems, Defense Systems, Construction Projects, Service Procurements like Security Services, Commercial Aviation Cargo Handling and many more. The scale of procurement ranges from hundreds of thousands of dollars to multi-billion dollar systems procurement.

The current version of eRFP uses multiple human evaluators to evaluate proposals against specific criteria and evaluation guiding rules. The System collects all the scores and comments from all evaluators, and provides a report based on scores and evaluator comments. It is a two-step process. First, proposals are shortlisted as acceptable or unacceptable. In the second step, top proposal is chosen for contract award. One substantial enhancement request from the clients of Sysoft eRFP is to assist the human evaluators with machine learning system that can guide human evaluators evaluate these proposals. Eventually it is desired that these human evaluators be replaced by the learning system enhancing the speed of evaluation and productivity. In addition, for some large and real complex procurement efforts, the machine learning system can be used to validate human agent based evaluation and act as "quick and simple" audit system to protect the public agencies and funds from fraud, corruption, unusual, and callas proposal evaluations by evaluators.

Every procurement entity evaluates proposals in slightly different manner. The criteria are also different depending on the RFP. Interestingly, industry norm is to provide some rule based objective guidance rules and recommendations for evaluators to evaluate these proposals. The rules can be as simple as "look for the word 'Neurogenesis' and see how many times used." As mentioned, these rules differ from one procurement agency to another and sometimes are dependent on Federal/State/local laws and internationally country specific laws. This is the reason why the learning system may be specific to a procurement entity. Eventually it may be possible to build a software configurator for tuning the specific attributes of the learning system to suit different agencies and RFP types in the future. This will make the system universal for all procurement usage.

Evaluating proposals from proposers is challenging for many reasons. Evaluators have to balance objective evaluation decisions against project usability objectives, specific criteria set by the procurement agency, and account for the increasingly diverse range of alternative technical approaches available in response to RFPs [1]. A potential solution to this problem is intelligent proposal evaluation support tools that can assist in proposal evaluation. These tools should be able to validate manual evaluations by human agents. However, in order to achieve this goal, such systems must have the capacity to represent *proposal* *evaluation attributes in a computable form*. eRFP proposal evaluation system should be able to represent human evaluation judgments in terms of variables in the evaluation space. Using machine-extracted proposal evaluation attributes from text-based features and evaluation judgments of 16 human evaluators on 52 proposals, we present preliminary results based on established methodologies that suggest that such representation may be possible [2]. This is the first step towards realization of a learning systems based automated proposal evaluation system.

In the remainder of this paper we summarize prior work in the area of statistical learning of "criteria based appeal" for proposals. We then describe the process of data collection and an experiment to investigate the possibility of representing proposal evaluation attributes for human readable textual proposals in a computable form. We conclude with a discussion of our findings and recommendations for future work in this technical area.

PREVIOUS WORK

Previous research has demonstrated that automatic text summarization based on word clusters and ranking algorithms allows one to adapt summaries to the user needs and to the corpus characteristics [3]. A review of the theory and methods of document classification and text mining, focusing on the existing literature shows possibilities of computable representations of human readable proposals in text[4]. The authors of the aforementioned work suggest that it is possible to develop computer algorithms that can deliver 'quick and dirty evaluations' of subjective aspects of any text based proposal. However, neither of these papers conclusively proves that proposal evaluation attributes can be represented in computable form.

We extend this previous research in a few significant and novel ways. First, we learn statistical models for predicting proposal evaluation attributes in terms of *text-based* features which have already been employed in text mining applications. These text-based features describe a proposal's use of fundamental corpus statistics: Word frequencies, co-occurrences, measure of point-wise mutual information for dependences between words, and word co-occurrences at shorter distances with word order [5]. They are informed by procurement agency's evaluation guidelines and evaluation subject matter experts (SMEs). Next, we employ performance evaluation to demonstrate that it may be possible for text-based feature models to learn and represent proposal evaluation attributes in a computable form.

COMPUTATION AND DATA ACEESS PLATFORM

The biggest challenge has been preparing the data for learning and testing. One specific procurement agency was selected for developing and testing prototype machine learning system that can evaluate the proposals based on responses from vendors and proposal evaluation guidelines set by the procurement agency. The responses are in free text format (unstructured data) but the evaluation guidelines are quite structured and objective. Sysoft eRFP is capable of capturing data on criteria, vendor responses, evaluation scores, and evaluation comments from evaluators who never evaluated these RFPs [6].

Word frequencies	Co-occurrences
Measure of point-wise mutual information for dependences between words	Word co-occurrences at shorter distances with word order

Figure 1. Our statistical models to predict proposal evaluation appeal were built using text-based features. These text based features are informed by interviews with subject matter expert evaluators, and evaluation guidelines from the procurement agency.

Sysoft eRFP works with Oracle and SQL Server. The database was segmented and replicated into MS Access with more than two hundred Tables. Matlab computing environment was connected to the database through Matlab JDBC/ODBC Relational Database connector. The data was readily available to Matlab computing environment through Structured Query Language commands to the Database from Matlab code. This made the computing and testing environment very compact and productive in a laptop running Matlab with MS Access.

EXPERIMENT

We conducted an experiment to investigate whether *text-based* feature models can reliably learn and predict proposal evaluation attributes on 52 proposals from two domains of proposal sets *(RFP1* and *RFP2).* RFP1 was "Procurements of Telephone Systems for Special Situations." RFP2 was "Procurement of Genome Analyzers for Hospital patients." Both RFPs belong to the same procurement agency.
Specifically, we formulated the following two hypotheses:

H.1 *text-based* feature models can reliably learn proposal evaluation attributes for a set of evaluations in the *RFP1 domain*
H.2 *text-based* feature models that are learned on the *RFP1 domain* can make reliable predictions on the *RFP2 domain.*

Proposal evaluation judgment refers to the objective evaluation opinion collected from the evaluators. Proposal evaluation attributes refer to the phenomenon that we are trying to learn and represent in a computable form.

Collecting criteria based appeal judgments

Data set of proposals

RFP1 received 30 proposals from various vendors. RFP 2 received 22 proposals from some other vendors. The same 16 evaluators evaluated all the 52 proposals. The evaluation criteria for both proposals were: How technically competent is this proposal in a scale of 0 to 100, 0 being worst and 100 being best. Evaluation data was captured in Sysoft eRFP System. The proposals were submitted in text format by the vendors through the Sysoft eRFP online proposal submission system. The evaluators were able to access the proposals via the eRFP evaluation interface that they are familiar with [6,7]. The evaluators evaluated the proposals based on the selected criteria. Sysoft eRFP automatically rank ordered these proposals for each evaluator based on their scores. The process was repeated for the 20 proposals from RFP1 domain and then the 32 proposals from RFP2 domain.

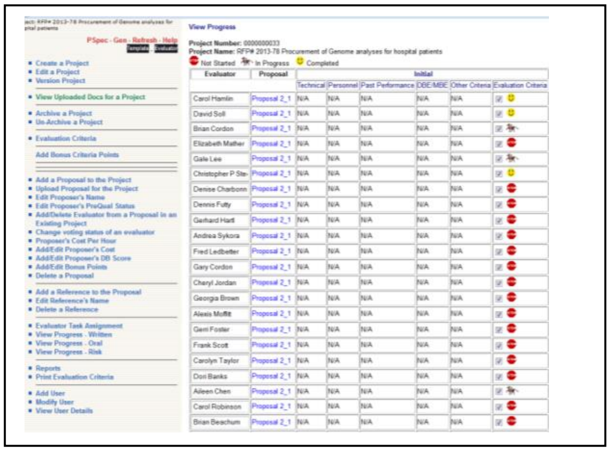
Proposal Evaluators

Sixteen evaluators from one procurement agency participated in this experiment. None of these evaluators had any affiliation with the vendors who submitted the proposals. Their contribution of effort and time was compensated by the volunteering procurement agency. None of these evaluators evaluated proposals of RFP1 or RFP2 before.

Procedure

We designed a two phase eRFP-based *rank ordering* experiment for measuring the proposal evaluation judgments. Before starting the experiment, evaluators were told to evaluate these proposals like any other evaluation project and were given standard time to evaluate these proposals. Evaluators were informed that they will evaluate first the 30 proposals for RFP1domain in the Sysoft eRFP evaluation interface and then repeat the same process for the next 22 proposals in RFP2domain.

Time limits were set for this experiment from previous similar projects completed in the same procurement agency. Evaluator used standard procurement agency guidelines.


ect: RFP# 2013-78 Procurement of Genome analyses for pital patients
PSpec - Gen - Refresh - Help
Template - Evaluator
Create a Project
Edit a Project
Version Project
View Uploaded Docs for a Project
Archive a Project
Un-Archive a Project
Evaluation Criteria
Add Bonus Criteria Points
Add a Proposal to the Project
Upload Proposal for the Project
Edit Proposer's Name
Edit Proposer's PreQual Status
Add/Delete Evaluator from a Proposal in an Existing Project
Change voting status of an evaluator
Proposer's Cost Per Hour
Add/Edit Proposer's Cost
Add/Edit Proposer's DB Score
Add/Edit Bonus Points
Delete a Proposal
Add a Reference to the Proposal
Edit Reference's Name
Delete a Reference
Evaluator Task Assignment
View Progress - Written
View Progress - Oral
View Progress - Risk
Reports
Print Evaluation Criteria
Add User
Modify User
View User Details

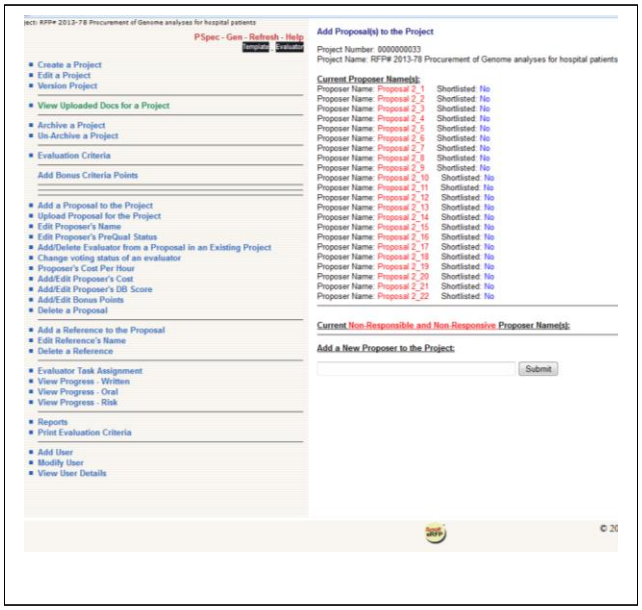
View Progress

Project Number: 0000000033
Project Name: RFP# 2013-78 Procurement of Genome analyses for hospital patients

🛑 Not Started 🏇 In Progress 🙂 Completed

Evaluator	Proposal	Initial					
		Technical	Personnel	Past Performance	DBE/MBE	Other Criteria	Evaluation Criteria
Carol Hamlin	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🙂
David Soll	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🙂
Brian Cordon	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🏇
Elizabeth Mather	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Gale Lee	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🏇
Christopher P Ste-	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🙂
Denise Charbonn	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Dennis Futty	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Gerhard Hart	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Andrea Sykora	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Fred Ledbetter	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Gary Cordon	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Cheryl Jordan	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Georgia Brown	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Alexis Moffit	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Gerri Foster	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Frank Scott	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Carolyn Taylor	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Dori Banks	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Aileen Chen	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🏇
Carol Robinson	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑
Brian Beachum	Proposal 2_1	N/A	N/A	N/A	N/A	N/A	☑ 🛑

Figure 2. Evaluation progress screen in eRFP


ect: RFP# 2013-78 Procurement of Genome analyses for hospital patients
PSpec - Gen - Refresh - Help
Template - Evaluator
Create a Project
Edit a Project
Version Project
View Uploaded Docs for a Project
Archive a Project
Un-Archive a Project
Evaluation Criteria
Add Bonus Criteria Points
Add a Proposal to the Project
Upload Proposal for the Project
Edit Proposer's Name
Edit Proposer's PreQual Status
Add/Delete Evaluator from a Proposal in an Existing Project
Change voting status of an evaluator
Proposer's Cost Per Hour
Add/Edit Proposer's Cost
Add/Edit Proposer's DB Score
Add/Edit Bonus Points
Delete a Proposal
Add a Reference to the Proposal
Edit Reference's Name
Delete a Reference
Evaluator Task Assignment
View Progress - Written
View Progress - Oral
View Progress - Risk
Reports
Print Evaluation Criteria
Add User
Modify User
View User Details
© 20

Add Proposal(s) to the Project

Project Number: 0000000033
Project Name: RFP# 2013-78 Procurement of Genome analyses for hospital patients

Current Proposer Name(s):

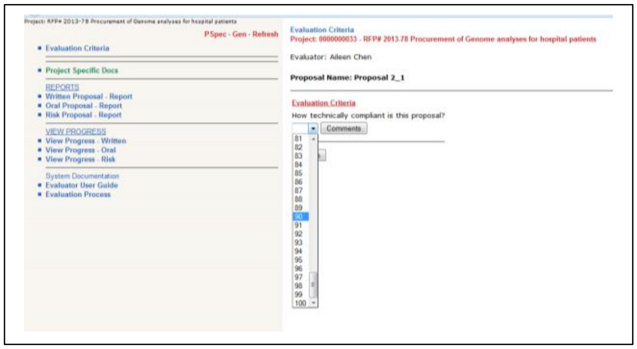
Proposer Name: Proposal 2_1	Shortlisted: No
Proposer Name: Proposal 2_2	Shortlisted: No
Proposer Name: Proposal 2_3	Shortlisted: No
Proposer Name: Proposal 2_4	Shortlisted: No
Proposer Name: Proposal 2_5	Shortlisted: No
Proposer Name: Proposal 2_6	Shortlisted: No
Proposer Name: Proposal 2_7	Shortlisted: No
Proposer Name: Proposal 2_8	Shortlisted: No
Proposer Name: Proposal 2_9	Shortlisted: No
Proposer Name: Proposal 2_10	Shortlisted: No
Proposer Name: Proposal 2_11	Shortlisted: No
Proposer Name: Proposal 2_12	Shortlisted: No
Proposer Name: Proposal 2_13	Shortlisted: No
Proposer Name: Proposal 2_14	Shortlisted: No
Proposer Name: Proposal 2_15	Shortlisted: No
Proposer Name: Proposal 2_16	Shortlisted: No
Proposer Name: Proposal 2_17	Shortlisted: No
Proposer Name: Proposal 2_18	Shortlisted: No
Proposer Name: Proposal 2_19	Shortlisted: No
Proposer Name: Proposal 2_20	Shortlisted: No
Proposer Name: Proposal 2_21	Shortlisted: No
Proposer Name: Proposal 2_22	Shortlisted: No

Current Non-Responsible and Non-Responsive Proposer Name(s):

Add a New Proposer to the Project:

Submit

Figure 3. Proposals entered in eRFP system



Evaluation Criteria
Project: 0000000033 - RFP# 2013.78 Procurement of Genome analyses for hospital patients

Evaluator: Aileen Chen

Proposal Name: Proposal 2_1

Evaluation Criteria
How technically compliant is this proposal?

Figure 4. Human evaluators evaluating and scoring proposals in eRFP evaluation interface

Evaluation progress was monitored through the Sysoft eRFP evaluation progress interface [6,7].

Analysis

1.Clustering criteria based appeal judgments at group level

We applied K-means on the median and standard deviation of rank positions of each proposal from the eRFP database to generate shortlisted clusters of "acceptable" and "unacceptable" proposals at the group level. The median rank of a proposal represents its degree of criteria based evaluation appeal at the group level. The standard deviation represents the level of agreement within the group about the criteria based evaluation appeal. Conceptually, a proposal with high median rank and relatively large standard deviation is considered an "unacceptable proposal with low group level agreement", while a proposal with low median rank and low standard deviation is considered an "acceptable proposal with high group level agreement". Established methodologies were used to perform this analysis [2].

2. Model building and statistical testing

After generating group-level (binary) labels, called shortlisting in procurement methodologies, we trained and tested three types of supervised learning models: Naïve Bays[8], Binary logistic regression[9], and Linear Kernel Support Vector Machines[10]. To test our hypothesis in Page 2, we evaluated each model using two measures of performance:

a. **Within-domain performance (H1)** : Leave One Out Cross Validation error on *RFP1 domain*

b. **Transfer performance (H2)**: Test error on *RFP2 domain* of model trained on *RFP1 domain.*

Statistical confidence testing was performed on each model's performance measures because of small number of examples in both domains. In other words, if and only if the following null hypothesis could be rejected using a 95% confidence level ($p<0.05$), a model is considered to have learned group-level proposal evaluation attributes with statistically significant performance:

H_0: Learned model error>= Model error of a model that makes decisions by randomly flipping a coin

We used the 'exact test for goodness of fit' to test for statistical significance [2].

RESULTS

Group-level 'shortlisting: acceptable and unacceptable' clustering

Computationally viable K-means with squared euclidian distance was run until convergence. Clustering in both domains was linearly separable. Squared distance made the computation efficient and feasible..

In the RFP1 domain, (n=22), acceptable (shortlisted) cluster contained 11 proposals while the unacceptable (not shortlisted) cluster contained 11 proposals. In the RFP2 domain, (n=30), acceptable (shortlisted) cluster contained 21 proposals while the unacceptable (not shortlisted) cluster contained 9 proposals. Figure 5 provides the detailed results:

Domain	Size of domain (n)	Shortlisted /not-shortlisted	Statistic	Median Rank	Std. Dev Rank
RFP1	**22**	**Shortlisted =11**	**μ**	**11.80**	**7.40**
			σ	**2.80**	**1.42**
		Not-shortlisted =11	**μ**	**21.2**	**9.45**
			σ	**2.54**	**1.68**
RFP2	**30**	**Shortlisted =21**	**μ**	**4.80**	**3.90**
			σ	**2.21**	**0.65**
		Not-shortlisted =9	**μ**	**11.68**	**4.92**
			σ	**1.91**	**0.70**

Figure 5: Group-level 'shortlisting: acceptable and unacceptable' clustering details

Supervised model evaluation

The Naïve Bayes model had a *within-domain performance* of 0.48 (False positive rate (FP) = 0.39, False negative rate (FN) = 0.57, not statistically significant (n.s.)) and *across-domain performance* of 0.58 (FP=0.31, FN=0.8, n.s.). The logistic regression model had a *within-domain performance* of 0.56 (False positive rate (FP) = 0.62, False negative rate (FN) = 0.50, n.s.) and *across-domain performance* of 0.65 (FP=0.0, FN=1, n.s.). The linear kernel SVM (KKT Tolerance =0.7, C=1) had a *within-domain performance* of 0.38 (FP= 0.38, FN = 0.38, p* < 0.05) and *across-domain performance* of 0.35 (FP= 0.34, FN = 0.35, p* < 0.05).

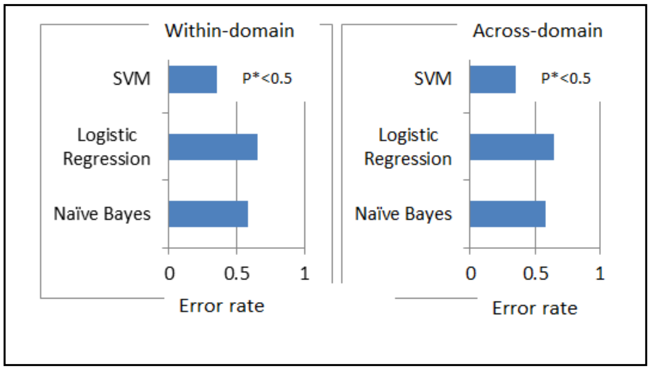


Figure 6: Graphs manifest performance of the three supervised learning models. Performance on the RFP1 domain is depicted by the 'Within-domain' graph. Performance of the models trained on RFP1 domain and tested on the RFP2 domain is depicted by the 'Across-domain' graph. Statistically significant performance is shown by the SVM model in 'Within-domain' and 'Across-domain'.

DISCUSSION

The standard deviation of cluster groups are comparable along the dimensions of degree of criteria based proposal evaluation appeal and group level agreement for both domains. Similar to other similar projects undertaken [2] this observation suggests that K-means provided a reliable estimate of group level proposal evaluation judgments, and furthermore that our analysis with binary labels was valid.

Our result shows that a linear kernel SVM has statistically significant performance on both within- and across- domain measures. These results verify our experimental hypotheses H1 and H2 (Page 2). Furthermore false positive and false negative rates also suggest that the model was finding statistical structure of group level proposal evaluation attributes in terms of *text-based* features.

OUTCOME

The results of this project has raised the interest level among the Sysoft eRFP decision makers to the extent that they have allocated funds to perform a twelve member Machine Learning support project starting January 2014. This will allow us to continue working on the project.

CONCLUSION AND FUTURE WORK

We aimed to explore whether it was possible to express proposal evaluation attributes in a computable form. The finding from our experiment suggests that such representation may be possible. The scope of this project was limited because of time limit and lack of resource allocation. The next step is to establish the findings from our experiment at a much larger scale across many different RFP domains and many procurement agencies across the country and internationally.

In future, we plan to engineer and find applicability of the following: Advanced Corpus Statistics, Statistical Text Mining Models, Geometrical Models, and Dimensionality Reduction and other Natural Language Processing (NLP) techniques to enhance our models[4,5,9,12,13,14]. Also, we may use document categorization with unsupervised clustering, supervised classification with Vector and Probability Space, content analysis with polarity estimation, property estimation, and property extraction [5,13].

We sincerely hope that our work lays the foundation for more extensive research in the computational techniques that can be used to research and implement intelligent evaluation support tools and eventually develop a completely automated evaluation system.

ACKNOWLEDGEMENTS

I would like to thank Arther Miller and Scott Palma for their insight into eRFP system and for co-advising this work at Sysoft Corporation. I would also like to thank Sheema Mirchandani for her guidance and support.

REFERENCES

1. Bennnan, P. Bringing RFP Evaluations Into The 21st Century, *Rockland County, NY, (2009).*
2. Lahiru G., J., Representing visual aesthetic tastes for wab pages in computable form, *Machine Learning Class Project*, Stanford University, (2011) Autumn.
3. Amini, M.., Usunier, N., and Gallinari, P., Automatic text summarization based on word clusters and rankings. *In proceedings of the In Proceedings of the 27 th European Conference on Information Retrieval*, (2005).
4. Khan, A., and Baharudin, B. A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information tTechnology 1*, 1 (2010), 183–195.
5. Branchs, R., Text Mining with Matlab, NY, Springer, (2013).
6. eRFP.sysoft.com: Sysoft eRFP website, (2011).
7. Adhikari, S.,: Intelligence: Learn, Infer, and Use Knowledge to Lower Cost, *Contract Management*, February, (2012), 66-70.
8. McCullagh, P., and Nelder, J. *Generalized linear models*. Chapman & Hall/CRC, (1989).
9. Mitchell, T. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill* (1997).
10. Shawe-Taylor, J., and Cristianini, N. An introduction to support vector machines and other kernel-based learning methods. *Cambridge University Press, UK* (2000).
11. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. Feature selection for svms. *Advances in neural information processing systems* (2001), 668–674.
12. Berry, M., W., Survey of Text Mining, Springer, NY(2003).
13. Jurafsky, D., Speech and Language Processing, *Upper Saddle River, Pearson*, NJ, (2009).
14. Sebastini, F., Machine Learning in Automated Text Categorization. *ACM Comput Surv*,(2002)34(1):1-4.