Learning Agents for Mining Domain-Specific Data

XXXXXX
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831
X-XXX-XXX-XXXX
XXXXX@ornl.gov

XXXXXX
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831
X-XXX-XXX-XXXX
XXXXX @ornl.gov

XXXXXX
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831
X-XXX-XXX-XXXX
XXXXX @ornl.gov

XXXXXX
Oak Ridge National Laboratory
P.O. Box 2008 MS 6085
Oak Ridge, TN 37831
X-XXX-XXX-XXXX
XXXXX @ornl.gov

ABSTRACT
In order to move beyond simple keyword searching for textual information retrieval, extensive effort is often used to build specialized ontologies and dictionaries that can significantly improve retrieval. Unfortunately, these efforts are generally not feasible for domain-specific data. What is needed is an automated means of learning the characteristic cue phrase patterns of a domain-specific language and using those learned patterns as a basis for automatically categorizing, clustering, or retrieving relevant data for the user. This work describes a multi-agent approach to learning domain-specific phrase patterns that utilizes a genetic algorithm (GA) as the learning method. Using the GA, the agents have the ability to learn phrase patterns from both successful and failed individuals in the GA population. This approach is applied to the analysis of mammography reports, which utilize a very specific language. The approach described here successfully learned phrase patterns for two distinct classes of mammography reports.

Categories and Subject Descriptors
I.2.6 [**Learning**]: analogies, concept learning, connectionism and neural nets, induction, knowledge acquisition, language acquisition, parameter learning.

H.3.3 [**Information Search and Retrieval**]: clustering, information filtering, query formulation, relevance feedback, retrieval models, search process, selection process.

General Terms
Algorithms, Design

Keywords
Learning agents, multi-agent system, genetic algorithm, information retrieval, maximum variation sampling, mammography reports.

1. INTRODUCTION
The state-of-the-practice in textual information retrieval continues to rely on keyword searching. Over the past decade, there have been numerous research efforts to move past the dependence on keywords. These efforts have included the use of various intelligent techniques such as intelligent agents, neural networks, genetic algorithms, and fuzzy logic. There have also been attempts to enhance the automation of categorizing and clustering data. In addition, the use of ontologies and term padding via dictionaries has also been heavily researched.

For general topics, these research efforts have shown some promise with automated categorization and clustering beginning to move into the state-of-the-practice. Unfortunately, these techniques are limited when handling domain-specific data. The foundational problem is a direct result of the complexity of language (e.g., language that is specific only to breast cancer and mammography reports). In order to move beyond simple keyword searching of domain-specific data, extensive effort is needed to build specialized ontologies and dictionaries that can significantly improve retrieval. Such effort requires significant time from multiple domain experts. In addition, there are no standards for the creation of such specialized ontologies or dictionaries.

Furthermore, the complexity of language introduces difficulties of semantics and syntax to the existing difficulties of term definitions and usage in domain-specific data. Typically, there are multiple ways of relaying the same meaning by using numerous combinations of terms and syntax. Unfortunately, there are no rules for using these various techniques and the development of specialized ontology would not help. As a result, meaningful information retrieval of domain specific data becomes nearly impossible with keyword searching as the user cannot possibly know all combinations of terms and syntax.

Therefore, what is needed is an automated means of learning the characteristic cue phrase patterns of a domain-specific language and using those learned patterns as a basis for automatically categorizing, clustering, or retrieving relevant data for the user. This paper describes preliminary work being performed to address the learning aspect of this approach. Section 2 will discuss the background of the domain-specific data being addressed by this work. Section 3 discusses some related works. Section 4 will describe the learning approach, while section 5 discusses results. Section 6 will discuss future work.

Cite as: Title, Author(s), *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Decker, Sichman, Sierra, and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. XXX-XXX.

2. BACKGROUND

This work focuses on the language domain of mammography reports. Mammography is the procedure of using low-dose X-rays to examine the human breast for the purposes of identifying breast cancer or other abnormalities. Currently, for each patient that undergoes a mammogram, there is at least one X-ray image and one textual report written by a radiologist. In the report, the radiologist describes the features or structures that they see or do not see in the image. Essentially, this report is meta-data that is written by a human subject matter expert about the image. In order to effectively train a computer-assisted detection (CAD) system, these reports could be mined and used as supplemental meta-data. Unfortunately, little work has been done to utilize and maximize the knowledge potential that exists in these reports.

In this preliminary study, unstructured mammography reports were used. These reports consisted of 9,000 patients studied over a 5-year period from 1993 to 1998. There are approximately 120,000 reports in this set. Each report generally consists of two sections. The first section describes what features the radiologist does or does not see in the image. The second section provides the radiologist's formal opinion as to whether or not there are suspicious features that may suggest malignancy (i.e., or the possibility that the patient has cancer). The set of reports also includes a number of reports that simply state that the patient canceled their appointment.

As discussed in [9] using a subset of this data, these reports vary in length. Some radiologists use more words than others when describing the same features. For example, in patients that do not exhibit any suspicious features, there are some reports that very simply state that there are no suspicious features. However, for the same patient in a different year, a different radiologist will provide a much more lengthy report that describes all of the suspicious features that did not exist.

To provide a better perspective of the challenge of mining these reports, consider the following question. Given a database of these reports, how does one retrieve those reports that represent abnormalities in the patient? In mammography, most patient reports will represent "normal" conditions in the patient. Consequently, the reports with "abnormal" conditions are rare (defining the difference between what is "normal" and "abnormal" is beyond the scope of this paper). Performing a cluster of these reports, most of the normal reports would cluster together while the abnormal reports would not form a cluster. This is because "abnormal" conditions tend to be very unique and very specific to a patient while "normal" conditions are much more generic and broad. Even if clustering provided value, clustering a very large database of these reports is exceptionally computationally expensive. Categorizing would be faster, however, the challenge remains of determining the appropriate categories, and even then, the abnormal reports may not categorize correctly.

The main problem of trying to find abnormal reports lies in the language that is used in mammograms. As discussed in [9], abnormal reports tend to have a richer vocabulary than normal reports. In addition, normal reports tend to have a higher number of "negation" phrases. These are phrases that begin with the word "no" such as in the phrase "no findings suggestive of malignancy." Consider the phrases shown in Table 1 and Table 2. These are the negation phrases that generally occur in normal reports and the ones shown here are samples of the variations that have been found. In the set of 120,000 reports used for this work, there were at least 286 variations of phrases for Table 1 and 1,231 variations of phrases for Table 2.

Table 1. Example phrases using "no" and "malignancy"

no malignancy
no mammographic features malignancy
no mammographic features suggestive of malignancy
no findings suggestive of malignancy
no significant radiographic features of malignancy
no radiographic findings suggestive of malignancy
no radiographic change suggestive of malignancy
no specific radiograpic features of malignancy
no mamographic evidence of malignancy

Table 2. Example phrases using "no" and "suspicious"

no mammographic finding suspicious
no strongly suspicious forms
no strongly suspicious features
no strongly suspicious masses
no radiographically suspicious masses
no developing suspicious clustered microcalcifications
no finding strongly suspicious
no new suspicious mass lesions
no suspicious linear branching forms

Consider the phrase shown in Table 3 and Table 4. These phrases tend to occur in abnormal reports (but may also occur in normal reports) and the ones shown here are samples of the variations that have been found. In the set of 120,000 reports used for this work, there were at least 52 variations of phrases for Table 3 and 691 variations of phrases for Table 4.

Table 3. Example phrases using "appearance" and "tissue"

appearance suggesting radiating strands of tissue
appearance suggestive of accessory breast tissue
appearance of normal glandular tissue
appearance of asymmetric fibroglandular tissue
appearance of fibroglandular tissue
appearance of glandular tissue
appearance of normal fibroglandular tissue
appearance of soft tissue densities bilaterally

Table 4. Example phrases using "additional" and "views"

additional views obtained today demonstrate variation
additional compression views
additional set of bilateral cc views
additional lateral views
additional mediolateral oblique views
additional mammographic views
additional bilateral craniocaudal views
additional bilateral lateral medial views

Considering the language variations shown previously, the task of retrieving those reports that represent abnormalities is daunting. The variations of terms and syntax create a combinatorial

explosion while, semantically, these combinations tend to mean the same thing.

The goal, then, is to develop an automated approach to learning the skip bigrams (or s-grams) of cue phrases in the domain-specific language that sufficiently characterize the reports such that information retrieval becomes both more accurate and simplistic while, at the same time, not being computationally intensive. S-grams are word pairs in their respective sentence order that allow for arbitrary gaps between the words [1][10]. The s-grams for Table 1 are the words "no" and "malignancy." This s-gram uniquely identifies a particular semantic in the language of mammography reports and enables the identification of all possible variations of such phrases. Higher-level patterns may then be formed from these s-grams. For example, the s-grams for Table 1 and Table 2 both imply that there are no abnormalities seen in the patient.

The work here describes a possible approach toward this goal of automatically learning s-grams that can provide meaningful retrieval on domain-specific data and the results achieved thus far.

3. RELATED WORKS

Further improvement in information retrieval techniques requires the continued development of algorithms whose basis lies in semantic extraction and representation. Information retrieval (IR) research began with simple representations of documents and the terms that they contained [16]. This research progressed into syntactic analysis such as co-occurrence, N-grams, part of speech analysis, and context-free grammars. Recently, IR research has continued to move toward a basis in semantics. Many of these approaches involve the use of ontologies, conceptual graphs, and language models such as described in [3][4][6][7][17]. Unfortunately, many of these approaches are either unable to scale, require significant effort on the part of subject matter experts, or do not handle domain specific data robustly. The work described here differs from these approaches in that it leverages computationally efficient, unsupervised learning of domain specific data in order to more effectively retrieve information. As a result, extensive ontologies are not needed, nor extensive effort on the part of a subject matter expert.

In [1], an unsupervised approach to identifying cue phrases is discussed. Cue phrases are formulaic patterns of phrases that have similar semantics but vary in syntactical and lexical ways. In [1], the authors use a lexical bootstrapping algorithm that relies on the use of "seed" phrases. While our work is addressing nearly the same problem, our work differs in that no seed phrases are needed, and that s-grams found for cue phrases using our approach are split into two classes.

Finally, the work described here is an extension of the work described in [9]. The goal for that work was "to find an ideal sample of mammography reports that represents the diversity without applying clustering techniques or without prior knowledge of the population categories." That approach successfully leveraged non-probabilistic sampling to retrieve documents that were as diverse as possible. The success of that approach was a direct result of the rich variations of the language used in mammography reports. The work described here continues that investigation of those rich variations.

4. APPROACH

As discussed in section 2 and in [9], mammography reports exhibit two characteristics. First, abnormal reports tend to have a wider variation in the language that is used. Consequently, these reports tend not to cluster with other reports. The second characteristic is that normal reports use more negation phrases than abnormal reports. It is these two characteristics that we seek to exploit in this approach.

To exploit the first characteristic, an enhancement of the maximum variation sampling technique [9] is developed. This technique is implemented via a genetic algorithm (MVS-GA) and is discussed in section 4.2 along with the enhancements. In addition, the work described here differs from [9] in that the MVS-GA is used to learn common phrase patterns among diverse documents and not explicitly for sampling. To exploit the second characteristic, the MVS-GA is augmented with a simple memory that stores the common phrase patterns of samples that failed to survive in the MVS-GA. This will be discussed in section 4.3.

Finally, intelligent software agents were created using this learning technique. Each agent analyzes different segments of the data set to learn the various language patterns. Currently, the agents work individually. However, future work will explore cooperative multi-agent learning. The multi-agent framework that was used for this work is described in the next section.

4.1 Multi-Agent Framework

All of the software agents in this system were developed using the Oak Ridge Mobile Agent Community (ORMAC). The ORMAC framework allows execution of mobile, distributed software agents, and establishes communication among them. The ORMAC framework provides a peer-to-peer communication and control topology (i.e., one agent can communicate with one or several other agents). This messaging approach provides the ability for communication that is encapsulated and asynchronous with a blackboard coordination model.

Messages are passed to a blackboard, and agents that are subscribed to the blackboard receive the messages. ORMAC enables an agent community to be quickly created using a set of machines with each machine executing the ORMAC agent host software. The ORMAC agent host software allows agents to migrate among machines. The ORMAC framework uses the Foundation for Intelligent Physical Agent (FIPA) compliant agent communication language (ACL) messages. Within the ORMAC community, each agent host is registered with a name server responsible for tracking where agents are currently being hosted [14].

The focus of this work is primarily on the learning algorithm described in the next section. However, future work will leverage the communication of agents to enable cooperative multi-agent learning.

4.2 Learning from Maximum Variation Sampling

Maximum variation sampling is a nonprobability-based sampling. This form of sampling is based on purposeful selection, rather than random selection. The advantage of this form of sampling is that it allows a doctor or radiologist to look at data that may not otherwise be visible via the random selection process. Since abnormal mammography reports are not as common as normal

ones, random sampling would make it difficult to find them. Within nonprobability-based sampling, there are several categories of sampling [8], one of which is maximum variation sampling (MVS) [8]. This particular sampling method seeks to identify a particular sample of data that will represent the diverse data points in a data set. In this case, the diverse data points will represent abnormal mammograms. According to [8], "This strategy for purposeful sampling aims at capturing and describing the central themes or principle outcomes that cut across a great deal of [data] variation." The MVS is naturally implemented as a genetic algorithm (MVS-GA).

Before applying a GA to the analysis of mammography reports, the reports must be prepared using standard information retrieval techniques. First, reports are processed by removing stop words and applying the Porter stemming algorithm [5][11][12]. Once this has been done, the articles are then transformed into a vector-space model (VSM) [13][16]. In a VSM, a frequency vector of word and phrase occurrences within each report can represent each report. Once vector-space models have been created, the GA can then be applied.

Two of the most critical components of implementing a GA are the encoding of the problem domain into the GA population and the fitness function to be used for evaluating individuals in the population. To encode the data for this particular problem domain, each individual in the population represents one sample of size *N*. Each individual consists of *N* genes where each gene represents one radiology report (each report is given a unique numeric identifier) in the sample. For example, if the sample size were 10, each individual would represent one possible sample and consist of 10 genes that represent 10 different reports. This representation is shown in the following figure.

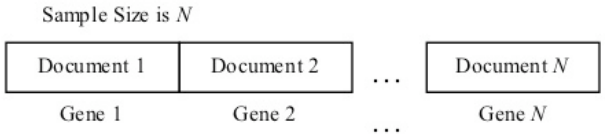

Figure 1. Genetic representation of each individual

The fitness function evaluates each individual according to some predefined set of constraints or goals. In this particular application, the goal for the fitness function was to achieve a sample that represents the maximum variation of the data set without applying clustering techniques or without prior knowledge of the population categories. To measure the variation (or diversity) of our samples, the summation of the similarity between the vector-space models of each document (or gene) in the sample is calculated as shown in the following equation.

In [9], the data was a test set that was selected by a human expert. For that work, the data was specifically chosen to test the ability of the MVS-GA. However, the data set for this current work utilizes a data set of approximately 120,000 reports. Within this data are numerous reports that simply state that the patient canceled their appointment. These reports are very small in length and are exceptionally distinct from all other reports (similarity values approaching zero). Unfortunately, the MVS-GA from [9] gravitates toward these cancellation reports as the best solution for a maximum variation sample.

In an effort to effectively characterize the phrase patterns of the mammography reports, it is necessary to examine reports that are longer in length, so that more language can be examined for patterns. In addition, abnormal reports tend to be longer in length than normal reports since the radiologist is describing the abnormalities in more detail. Consequently, the fitness function of the MVS-GA was enhanced to incorporate penalty functions as shown in equations 1 – 3.

$$Fitness(i) = \sum_{j=0}^{N} \sum_{k=j+1}^{N} \alpha_j + \beta_k + Similarity\,(Gene(i,j), Gene(i,k))$$

Equation 1. Revised MVS-GA Fitness Function

$$\alpha_j = e^{-\left(\frac{\|j\|}{100}\right)}$$

Equation 2. Penalty factor for document *j*

$$\beta_k = e^{-\left(\frac{\|k\|}{100}\right)}$$

Equation 3. Penalty factor for document *k*

In Equation 1, the Similarity function calculates the distance between the vector space models of gene *j* and *k* of the individual *i*. This distance value ranges between 0 and 1 with 1 indicating that the two reports are identical and 0 indicating that they are completely different in terms of the words used in that report. Therefore, in order to find a sample with the maximum variation, Equation 1 must be minimized (i.e., lower fitness values are better). In this fitness function, there will be $(N^2 - N) / 2$ comparisons for each sample to be evaluated.

The penalty functions are incorporated into the fitness function in order to penalize individuals in the MVS-GA based on the length of the documents they represent. Shorter documents receive higher penalties while longer documents receive much lower penalties. The penalty functions also return values that are between 0 and 1, inclusive. As a result of the penalty functions, the cancellation reports will receive the highest fitness values, while lengthy, abnormal reports will receive the lowest fitness values.

After the MVS-GA is executed, the end result is a best sample of mammography reports that are as diverse from each other as possible. Once this sample is achieved, then phrases are extracted from each document in the sample. For each phrase in the document, s-grams are extracted. Next, the s-grams are counted across the sample of documents. S-grams that are common across the sample will have higher frequency counts while s-grams with a frequency of 1 are uniquely identify a particular document in the sample. For this work, only those s-grams that are the most common in the best sample found are considered valuable. It is these s-grams that have the ability to uniquely retrieve abnormal documents from a large set.

4.3 Learning from Failures

Genetic algorithms (GA) are nature-inspired algorithms that mimic the natural selection process. The natural selection process is generically defined as survival of the fittest (i.e., only the most fit individuals for a given environment survive and reproduce offspring). During this process, the offspring are created from the best individuals; therefore, the population should continue to improve over several generations. It has been shown that canonical genetic algorithms converge to an optimal solution if the best individual remains in the population [15].

The primary intent of the GA is to converge toward an optimal solution. However, very little GA research, if any, has been performed that leverages knowledge gained from the individuals that failed to be selected and reproduce. In a typical GA, individuals that are not selected for reproduction are simply discarded.

For this work, the MVS-GA has been augmented to store the common s-grams of the failed individuals. This will enable answering questions such as what characteristic phrases make failed individuals inferior to successful individuals. After each generation, s-grams and their frequencies from each failed individual are extracted from each individual and stored in memory. After the MVS-GA has completed, the memory now contains the most common s-grams that caused individuals to fail in the GA. The end result is that the MVS-GA learns the s-grams for both abnormal and normal classes of reports.

5. RESULTS

The s-grams discovered by the learning agents on the data set are shown in Tables 5 and 6. Table 5 shows the top ten s-grams from the best solution obtained by the learning agents using MVS-GA. These s-grams tend to uniquely define abnormal reports. Many of these s-grams refer to procedures that are performed in the event that a suspicious feature in the patient was observed by the radiologist. For example, the patient may be asked to return with a few weeks for additional imaging such as an ultrasound and magnification imaging. In addition, patients with suspicious features may undergo biopsy, and in some cases, may also have a needle localization performed to enhance the biopsy procedure. Furthermore, since breast cancer often affect the lymph nodes, radiologist look for abnormalities relating to the lymph nodes as well. As can be seen in Table 5, the MVS-GA successfully learned key s-grams that would significantly enhance automated retrieval of abnormal reports.

Table 5. Top Ten s-grams from best solution obtained by MVS-GA

Rank	S-gram	Example Phrase	Number of Variations Observed
1	left & breast	left breast demonstrating apparent distortion	3640
2	core & biopsy	stereotactic guided core biopsy of microcalcifications	636
3	compression & views	additional bilateral anterior compression mlo views	762
4	spot & views	laterally exaggerated craniocaudal spot views	838
5	magnification & views	magnification views requested	648
6	spot & compression	mediolateral oblique spot compression views	1094
7	needle & localization	ultrasound-guided needle localization procedure	233
8	nodular & density	showing questionable increased nodular density	2701
9	lymph & node	atypically located intramammary lymph node	717
10	spot & magnification	breasts requiring spot magnification imaging	624

Table 6 show the top ten s-grams that begin with the “no” and were learned from the failed individuals in the MVS-GA. As discussed previously, most normal reports contain some form of a “negation” phrase. These phrases refer to the non-existence of a particular feature or condition in which the radiologist was searching. Abnormal reports may contain such negation phrases, however, abnormal reports tend to be more focused on the abnormalities that were found and not the abnormalities that were not found. Consequently, the learning agents using MVS-GA successfully learned from the failed samples the key s-grams of normal reports.

Table 6. Top Ten s-grams with the word "no"

Rank	S-gram	Example Phrase	Number of Variations Observed
1	no & suspicious	no finding strongly suspicious	1231
2	no & masses	no new focal masses	365
3	no & focal	no dominant focal lesion	210
4	no & evidence	no evidence of cyst	716
5	no & specific	no specific palpable abnormality detected	187
6	no & findings	no current physical findings	308
7	no & mass	no development of abnormal dominant mass	534
8	no & mammographic	no persisting mammographic abnormalities	390
9	no & radiographic	no radiographic lesions seen	285
10	no & calcifications	no clear cut clustered punctate calcifications	138

One of the most significant aspects of these results is that the learning agents did not require any specialized ontology or dictionary or feedback from a subject matter expert. The learning agents utilized an unsupervised, domain independent learning algorithm to achieve these results. Now that the agents have learned the s-grams, the agents can then begin retrieval of relevant documents. Future work will examine the retrieval quality of this approach.

6. FUTURE WORK

While the work described here focuses primarily on the learning aspect of mining domain-specific data, there are many avenues for future research. First, this work uniquely identified s-grams that defined two classes of mammography reports (abnormal and normal). Other data sets may have more than two classes of data, and so future work will investigate the expansion of this approach to identify *n* classes of data. Secondly, the work focused on a single learning algorithm for an intelligent software agent. However, intelligent agents have additional capabilities that can be utilized. To further enhance the learning capability and domain flexibility, future work will investigate cooperative agent learning to enhance this approach. Finally, the current approach

used a very rudimentary memory. A more advanced cognitive memory model will be explored in the future.

7. ACKNOWLEDGMENTS

Our thanks to Robert M. Nishikawa, Ph.D., Department of Radiology, University of Chicago for providing the large dataset of unstructured mammography reports, from which the test subset was chosen.

8. REFERENCES

[1] Abdalla, R. M., and Teufel, S. 2006. A bootstrapping approach to unsupervised detection of cue phrase variants. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (Sydney, Australia). COLING 2006. ACM Press, New York, NY, 2061-2064. DOI= http://dx.doi.org/10.3115/1220175.1220291

[2] Cheng, W., Greaves, C. and Warren, M. 2006. From n-gram to skipgram to concgram. International Journal of Corpus Linguistics 11/4: 411–33.

[3] Dridi, O.; Ben Ahmed, M., "Building an Ontology-Based Framework For Semantic Information Retrieval: Application To Breast Cancer," Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on , pp.1-6, 7-11 April 2008. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4530001&isnumber=4529902

[4] Duh, K., and Kirchhoff, K. 2004. Automatic learning of language model structure. In Proceedings of the 20th International Conference on Computational Linguistics (Geneva, Switzerland). COLING 2004. ACM Press, New York, NY, 2061-2064. DOI= http://dx.doi.org/10.3115/1220355.1220377

[5] Fox, C. 1992. “Lexical analysis and stoplists.” In Information Retrieval: Data Structures and Algorithms (ed. W.B. Frakes and R. Baeza-Yates), Englewood Cliffs, NJ: Prentice Hall.

[6] Jing-Yan Wang; Zhen Zhu, "Framework of multi-agent information retrieval system based on ontology and its application," Machine Learning and Cybernetics, 2008 International Conference on , pp.1615-1620, 12-15 July 2008. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4620664&isnumber=4620590

[7] Kai Kang; Kunhui Lin; Changle Zhou; Feng Guo, "Domain-Specific Information Retrieval Based on Improved Language Model," Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on , pp.374-378, 24-27 Aug. 2007. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4406104&isnumber=4406026

[8] Patton, M.Q. 1990. Qualitative Evaluation and Research Methods, Second Edition. Newbury Park, CA: Sage Publications, Inc.

[9] Patton, R.M., Beckerman, B., and Potok, T.E. 2008. Analysis of mammography reports using maximum variation sampling. In Proceedings of the 2008 GECCO conference companion on Genetic and Evolutionary Computation (Atlanta, GA). GECCO 2008. ACM Press, New York, NY, 2061-2064. DOI= http://doi.acm.org/10.1145/1388969.1389022

[10] Pirkola, A, Keskustalo, H., Leppänen, E., Känsälä, A.and Järvelin, K. 2002. "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants." Information Research, 7(2) [Available at http://InformationR.net/ir/7-2/paper126.html]

[11] Porter, M. 1980. “An algorithm for suffix stripping.” Program vol. 14, pp. 130-137.

[12] Porter Stemming Algorithm. Current Feb. 5, 2009. http://www.tartarus.org/~martin/PorterStemmer/

[13] Raghavan, V.V., and Wong, S.K.M. 1986. “A critical analysis of vector space model for information retrieval.” Journal of the American Society for Information Science, Vol.37 (5), p. 279-87.

[14] Reed, J.W., Potok, T. E., and Patton, R.M. 2004. "A multi-agent system for distributed cluster analysis," in Proceedings of Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems (SELMAS'04) Workshop in conjunction with the 26th International Conference on Software Engineering Edinburgh, Scotland, UK: IEE, pp. 152-5.

[15] Rudolph, G., "Convergence analysis of canonical genetic algorithms," Neural Networks, IEEE Transactions on , vol.5, no.1, pp.96-101, Jan 1994. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=265964&isnumber=6672

[16] Salton, G. 1983. Introduction to Modern Information Retrieval. McGraw-Hill.

[17] Siddiqui, T.J., "Integrating notion of agency and semantics in information retrieval: an intelligent multi-agent model," Intelligent Systems Design and Applications, 2005. ISDA '05. Proceedings. 5th International Conference on , pp. 160-165, 8-10 Sept. 2005. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1578778&isnumber=33356