McGlashen, M. (2002). Classification in the Physical Sciences. 13th ASIS SIG/CR Classification Research Workshop, 59-62.
doi:10.7152/acro.v13i1.13837

Classification in the Physical Sciences

Michael L. McGlashen & Anne Rogers
Business Intelligence Center
The Dow Chemical Company
mlmcglashen@dow.com; aerogers@dow.com

1. INTRODUCTION

Over the past decade, text classification researchers have published numerous studies (Hersh, Buckley, Leone & Hickman, 1994; Joachims, 1999; Lam & Ho, 1998; Lam, Ruiz & Srinivasan, 1999; Lewis, Schapire, Callan & Papka, 1996; McCallum & Nigam, 1998; McCallum, Rosenfeld, Mitchell & Ng, 1998; Yang, 1996; Yang, 1999; Yang & Pedersen, 1997) of supervised or unsupervised machine-learning techniques applied primarily to experimental data sets, such as Reuters-21578 or OHSUMED. In the commercial arena, vendors now promise turn-key solutions that will aggregate content from multiple, disparate sources, organize it into a body of useful information, and send just the right information to the right person, at the right time. While such tools are of enormous potential value to R&D-intensive science and technology organizations like The Dow Chemical Company, translating academic research and vendor hyperbole into operational success remains a formidable challenge.

Part of this difficulty stems from differences between the content and formatting of test collections used in experimental research settings and the documents typically found in a "real-world" science and technology environment. While collections of Reuters news articles or Medline abstracts are ideal vehicles for algorithm development, their applicability as surrogates for typical R&D documents is questionable. For example, the Medline abstracts used by Hersh et al. (1994) and others are, by definition, short *summaries* of a complete medical document; they are not full documents of themselves. They are relatively homogeneous, both in content and format, and information-dense, due to multiple human review steps, including: 1. Authors typically craft a title and abstract to convey only the main concepts described in the full document; 2. Peer reviewers in the journal publication process may also refine the title and abstract, if necessary; 3. Editorial selection of appropriate journals for indexing in the Medline database ensures that published abstracts are all related to Health and Medicine, and 4. Most researchers select a subset of these abstracts (e.g., Heart Disease) for detailed study.

In contrast, typical R&D collections generally contain complete documents in multiple (or without) formats, highly diverse subject matter and an uneven distribution of information quality. A study by McCallum et al. (1998) which reports the classification of Yahoo! Science records is probably the closest available approximation to true R&D content, however, the accuracy of this method (~39%) is too low to be useful in an operational setting. Classification of complete Physical Science documents remains a challenge due to several factors, including:

1. Nomenclature
Chemical and Biological names frequently comprise several words that combine to form a single name (e.g., sodium aluminum silicate). Other complicating factors include the existence of

ISSN: 2324-9773

McGlashen, M. (2002). Classification in the Physical Sciences. 13th ASIS SIG/CR Classification Research Workshop, 59-62.
doi:10.7152/acro.v13i1.13837

multiple synonyms for a single species, several nomenclature conventions, and non-standard (trivial) names created by the author. Multiple product names applied to various grades or formulations of the same chemical substance provide additional complexity. Accordingly, conventional algorithms that treat a document as a "bag of words" are likely sub-optimal for classifying Physical Science documents.

2. Long Documents with Multiple Sections

Physical Science documents are generally much longer than news reports and are usually comprised of several sections, including an Abstract, Background, Experimental, Detailed Discussion, Conclusions, and Bibliography. Formatting varies widely for different document types, however, classification algorithms that utilize this general document structure will provide advantages over existing methods. For example, although topics described in the background section may be mentioned frequently in the text, they are often only tangentially related to the main thrust of the article. Ideally, classification algorithms should be sensitive to the existence of document structure, but not rigidly confined by it.

3. Intended Audience

Another challenge relates to the differences in writing styles employed by authors in the Physical Sciences versus news articles. News article authors make few assumptions about the prior knowledge of the reader. In contrast, Physical Science documents are typically written for experts in the field who understand the subject domain. Concepts embodied in a Physical Science document may also infer relationships to related broader or narrower concepts that are not explicitly described. Classification methods which can reference the expertise of a given author by association with information external to the document in question (e.g., through the use of thesauri, taxonomies, cited or citing references, or other database-accessible information, such as chemical structures or reactions) would be advantageous. Aronson et al and others have previously demonstrated some success in this regard.

2. RECENT PROGRESS

In collaboration with our technology partners, The Dow Chemical Company has recently developed and deployed a range of classification tools which begin to address some of the aforementioned obstacles. As the result of a merger with Union Carbide Corporation in 2001, Dow acquired 300,000 detailed reports describing chemical and biological technologies and their applications. In order to maximize the value of its new intellectual capital, Dow initiated a project with ClearForest™ and IFI™ to provide detailed subject indexing for these reports by the end of 2002. In the course of that work, we have developed several novel Information Extraction and Text Classification techniques and have identified a number of promising opportunities for further research. Some of the project highlights include:

1. Machine-assisted identification of new-to-Dow (Union Carbide) chemical substances and addition of the same into the Dow Registry system (an index of chemical substances, their unique identifiers, and synonyms).
2. Selection of the "most valuable" documents for subsequent processing based on the following characteristics: source, author(s), format, publishing organization(s), document

ISSN: 2324-9773

McGlashen, M. (2002). Classification in the Physical Sciences. 13th ASIS SIG/CR Classification Research Workshop, 59-62. doi:10.7152/acro.v13i1.13837

type (e.g., Summary/Review vs. Technical Report vs. Trip Report), and the relevance of key document concepts to corporate business objectives, as determined by Dow experts.

3. Development of novel information extraction techniques which can identify chemical substances appearing in any document and provide a fuzzy "best match" to the Dow Registry. Although several prior studies have demonstrated the advantages of leveraging pre-existing fixed taxonomies and thesauri (Aronson, Rindflesch & Brown), our method uses a dynamically generated variant of the substance name for matching purposes. Chemical substance "recall" (recognition and correct matching) for this algorithm approaches 80%, as judged by a team of human reviewers.
4. Support of human review and associated work processes across a wide-area-network.
5. Early progress on the use of rotating, multiple orthogonal taxonomies for classifying physical science reports.

3. THE NEED FOR BENCHMARKS

While we have tested several methods for improving the classification of Physical Science documents, a new TREC repository is needed to adequately model the content and format complexity of typical science and technology environments. This collection would support the development of new text classification algorithms and provide a useful benchmark against which to evaluate competing vendor solutions.

. Ideally, this collection would comprise a wide array of document types and formats including published journal articles, corporate "gray" literature, and patents, and would cover a broad spectrum of technology subjects. Commercial database abstracts and indexing, along with the fulltext journal articles and patents corresponding to those abstracts, could be leveraged to provide ready-made training and testing sets. Additional source material could be gathered from academic and corporate participants.

Development of such a collection would likely require a collaborative effort between Government, Academic and Corporate partners, commercial publishers (e.g., Elsevier, Wiley, or the American Chemical Society) and commercial database producers (e.g., Derwent, Chemical Abstracts Service, or Elsevier, to name a few). Interested participants should be encouraged to identify and solicit participation by journal publishers and database producers in support of the project.

With the aforementioned benchmarks and appropriate metrics, the classification community could play a strategic role in helping science and technology organizations to manage their growing information volume. Benchmarks and best practices for the application of classification tools in an operational science and technology environment would be of enormous economic benefit.

ISSN: 2324-9773

McGlashen, M. (2002). Classification in the Physical Sciences. 13th ASIS SIG/CR Classification Research Workshop, 59-62. doi:10.7152/acro.v13i1.13837

REFERENCES

Aronson, A.R., Rindflesch, T.C. & Brown, A.C. (n.d.) *Exploiting a Large Thesaurus for Information Retrieval*, retrieved from http://skr.nlm.nih.gov/papers/references/riao94.final.pdf.

Hersh W., Buckley C., Leone T.J. & Hickman, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Croft, B. and Van Rijsbergen C.J. (Eds.) *Proceedings of the 17th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, July 1994*, pp. 192–201.

Joachims, T. (1999). Estimating the generalization performance of a SVM efficiently. *Technical report LS-8 Report 25, Dec. 1999.* Universit"at Dortmund, Dortmund.

Lam, W. & Ho, C.Y. (1998). Using a generalized instance set for automatic text categorization. In: *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 81–89.

Lam,W., Ruiz, M.E. & Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *IEEE Transactions on Knowledge and Data Engineering, 11*(6): pp. 865–879.

Lewis, D.D., Schapire, R.E., Callan, J.P. & Papka, R. (1996). Training algorithms for linear text classifiers. In: *Proceedings of the 19th International ACMSIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, July 1996*, pp. 298–303.

McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In: *Learning for Text Categorization: Papers from the 1998 Workshop. AAAI Technical Report WS-98-05, San Francisco, CA, July 1998.* Menlo Park, CA: AAAI Press, pp. 41–48.

McCallum, A., Rosenfeld, R., Mitchell, T. & Ng, A.Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. In: *Proceedings of the 15th International Conference on Machine Learning, AAA1, July 1998*. San Francisco, CA: Morgan Kaufman.

Yang, Y. (1996). An evaluation of statistical approaches to MEDLINE indexing. In: *Proceedings of the American Medical Informatic Association (AMIA)*, pp. 358–362.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*(1):69–90.

Yang, Y. & Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*. San Francisco, CA: Morgan Kaufmann.

ISSN: 2324-9773