December 2002

# A COMBINATORIAL APPROACH TO KNOWLEDGE DISCOVERY: THE INTEGRATION OF HUMAN AND MACHINE GENERATED KNOWLEDGE

Gondy Leroy
*University of Arizona*

# A COMBINATORIAL APPROACH TO KNOWLEDGE DISCOVERY: THE INTEGRATION OF HUMAN AND MACHINE GENERATED KNOWLEDGE

**Gondy Leroy**
University of Arizona
gleroy@bpa.arizona.edu

**Abstract**

*Human and machine generated knowledge have different strengths and weaknesses. Human knowledge is precise but has often limited coverage. Machine generated knowledge is less precise but can cover more ground efficiently. Knowledge based systems should tightly integrate both. Unfortunately, only few ways of combining human and machine generated knowledge will be practical and efficient. Three such knowledge integration approaches are developed and tested: human judgments to guide probabilistic and evolutionary information retrieval techniques, ontologies to provide the semantic context for an automatically created thesaurus, and ontologies to augment natural language processing. Human generated, precise, domain-specific ontologies were found to be well suited for integration with both machine learning algorithms and natural language processing for knowledge discovery. These conclusions are applied in the development of GeneScene, a knowledge based system being developed for biomedicine. GeneScene relies on a novel natural language processing technique: a 'function word'-based parser. This parser is integrated with medical ontologies to extract biomedical pathway information from text.*

**Keywords:** knowledge discovery, knowledge management, ontology, machine learning, natural language processing

## Introduction

The desire for knowledgeable and useful computer programs goes back to the 70s when the first expert systems were built. Although many expert systems are in use today, the necessary knowledge for new systems is hard to acquire and becomes easily outdated in rapidly growing fields. Machine learning was therefore attempted and aimed initially to supply knowledge to expert systems by automating the knowledge discovery process from existing digital information sources through exploitation of regularities in the data (Langley and Simon, 1995). Machine learning has since been expanded and is today, for example, the main component of data and text mining. Unfortunately, automated knowledge discovery is seldom sufficiently precise or immediately fit for human consumption. Especially in the biomedical field, one of the fastest growing research fields, researchers require more precise knowledge representation than existing software can provide.

To effectively and efficiently develop knowledge repositories, human knowledge and machine generated knowledge need be combined in a manner that maximizes the potential of each. The strength of human knowledge lies in its precision. The strength of machine generated knowledge lies in the efficiency of machine learning techniques to extract knowledge from large information collections without human intervention. However, few methods of combining human and machine generated knowledge allow effective development of new, easily accessible knowledge repositories. In the following sections, test case studies to combine human and machine generated knowledge are discussed. The lessons learned are used in the development of GeneScene, a knowledge repository for biomedical researchers that will combine literature resources and experimental data .

## Knowledge Generation

### *Human Knowledge Generation*

Knowledge differs from both data and information. Data are sets of facts, often represented by numbers, and information informs us about the data. "Information is data that makes a difference" (Davenport and Prusak, 1998). Knowledge is information that is mentally processed by people and becomes part of their representation of the world. The context is the person's knowledge base used for interpreting the information and the nature of the problem he/she faces. Knowledge becomes information again once it is articulated (Alavi and Leidner, 2001) and removed from its context.

In the past, developers of information systems gathered human knowledge painstakingly by building heuristics through interaction with domain experts. Although very valuable, the knowledge is mostly tailored for one application. Ontology building provides a more systematic approach to gathering knowledge for a specific domain and makes it easier to share the knowledge with others (Gruninger and Lee, 2002). An ontology is an explicit representation of a conceptualization (Grüninger and Uschold, 1996), or human knowledge expressed by a representational vocabulary. The available ontologies cover diverse topics and depending on their intended use, they can have different levels of abstraction. Some are defined in natural language; others use a strict and formal language. Two well known ontologies are WordNet, developed at Princeton University (http://www.cogsci.princeton.edu/~wn/), and the Unified Medical Language Systems (UMLS), developed by the National Library of Medicine (http://umlsks.nlm.nih.gov/).

### *Machine Knowledge Generation*

Knowledge can be automatically generated from textual and other data. Natural language processing (NLP) tools are required to preprocess free text and extract the relevant parts. A regular expression can describe specific string sequences and can be implemented with finite state automata (FSA), which are extremely efficient. More sophisticated tools, i.e., full parsers, can provide the syntactic structure of complete or parts of sentences depending on the need of the user. They can be based on context free grammars and can use different search algorithms to build a fitting parse tree. They often combine probabilistic information with context free grammars or with lexical information to find a sentence's parse tree.

There are several probabilistic techniques and machine learning paradigms (Chen, 1995) useful for text mining. The probabilistic approach is based on the probabilities of words in text. It matches queries to documents based on probabilities of relevance, or builds probability-based networks of documents and queries. Symbolic learning techniques represent knowledge symbolically, e.g. as rules or hierarchies. Neural networks, on the other hand, retain their knowledge in their nodes and connections. Genetic algorithms use an evolutionary paradigm where a population of individuals progresses over different generations to produce fitter individuals that represent the best state, value, or answer to a problem.

## Integrating Human and Machine Generated Knowledge

### *Human Judgments to Augment Probabilistic and Evolutionary Techniques*

Relevance feedback and genetic algorithms can help casual users search the Internet with a search engine. Search engines themselves do not provide sophisticated methods to help casual Internet users access relevant web pages. The users' domain knowledge, experience, and success searching the Internet are varied, and frequently the search results are poor. The goal of this study was to discover if machine learning techniques can deduce the user interest based on the available user judgments. Previous research on automatic query expansion reveals that both relevance feedback and genetic algorithms based on user feedback are good candidates to improve search results. With many applications where the texts are peer-reviewed documents and a large amount of feedback is available, both algorithms produce substantial increases in performance (Chen, et al., 1998; Fan, et al., 2000; Kraft, et al., 1994; Salton and Buckley, 1990).

Two algorithms draw on implicit user feedback from a user to expand that user's query with additional keywords (Leroy, et al., submitted), as can be seen in Figure 1. Each algorithm can expand the query with keywords from implicit positive feedback (positive expansion), or from implicit negative feedback (negative expansion) and then submits these to the Google search engine (http://www.google.com). In the first case, keywords are added with a Boolean "and," in the second with a Boolean "not." The hypotheses were that the genetic algorithm would increase performance more than relevance feedback, that expert users would

guide the search with good keywords and would therefore benefit most from the additional filtering with negative expansion, and that novice users would need help formulating a good query and would therefore benefit most from positive query expansion.

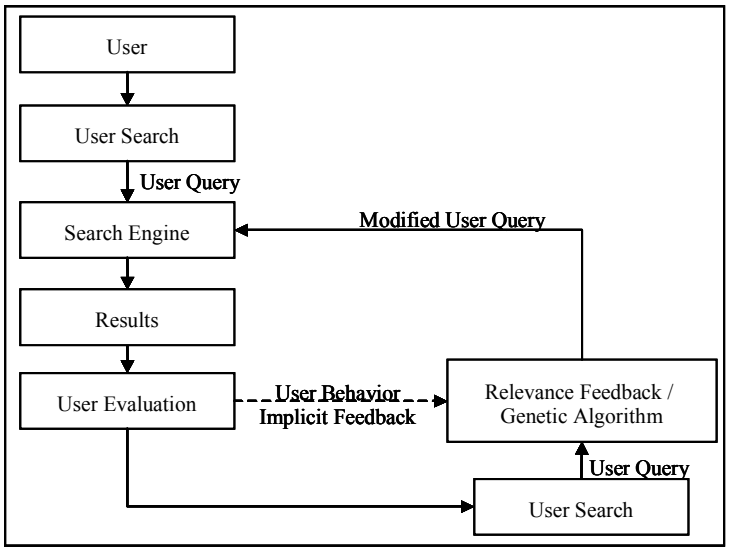


**Figure 1. Human Judgment to Guide Relevance Feedback and Genetic Algorithms**

This study was designed as a within-subjects study so that each subject participated in all experimental conditions, including a baseline condition without algorithms. Twenty-six subjects solved five questions randomly assigned to the conditions. Experts compiled a gold standard against which the user data was compared. The average F-measure over all conditions for each subject was used to divide the group post-hoc into three achievement groups: high, middle, and low achievers. For the low achievers, precision was significantly higher with the genetic algorithm and negative expansion ($p < .05$) and there was a strong trend ($p = .117$) for improved recall under the same condition. In contrast, for the high achievers, there was a strong trend that precision was higher with relevance feedback than with the genetic algorithm ($p = .09$). These results indicate that novice users are helped by negative expansion instead of positive expansion contradictory to our expectations. The feedback loop did not provide sufficiently high-quality information to help other users. Some users did not provide enough evaluation of the results or provided inconsistent evaluation of the results, leading to contradictions in the expanded queries.

On-the-fly deduction of human knowledge and judgment is not a sufficient base to guide machine learning algorithms to find previously unknown knowledge online. Although novices were helped by the evolutionary techniques, the quality of the feedback is often inconsistent and insufficient to provide precise guidance for automated knowledge discovery. A broader and more systematic knowledge representation is required.

### *Ontologies to Augment a Probabilistic Network*

A lack of vocabulary compatibility between user and the information system frequently impairs searches. This may be due to lack of or differences in expertise, and can often be remedied with an appropriate thesaurus. Concept Space is an automatically created medical thesaurus based on a probabilistic network developed for this purpose (Chen and Lynch, 1992). However, because the thesaurus terms are insensitive to a user ‘s query context, a semantic parsing algorithm based on WordNet and the UMLS ontologies, especially the Semantic Net, was designed and built to augment the thesaurus to increase the precision of the thesaurus terms and the compatibility of the terms with user queries (Leroy and Chen, 2001), see Figure 2. Completely automated tools that make use of the Semantic Net are scarce because of its structure: it contains concepts, the semantic types they belong to, and the semantic relations between the types. Difficulties can arise when automated tools wish to make use of these relations because the relations exist between semantic types, but not necessarily between the concepts that belong to those types. For example, the semantic type “*Medical Devices*” has a “*treat*” relation with the semantic type “*Sign and Symptom*.” However, “*Bone screws*” (Medical Device) do not treat “*nausea*” (Sign or Symptom). Consequently, assumptions based on top-down relations, will almost always be incorrect.

The semantic parsing algorithm, maps user queries and related thesaurus terms onto the ontology structure. The algorithm first establishes the context for the user query by mapping the terms to the Semantic Net. It then retrieves terms related to the user query from the thesaurus. Terms that do not fit the semantic context established for the query are excluded; the others are presented to the user. By limiting the thesaurus terms in this manner, i.e. bottom-up mapping of terms onto the ontology structure, the terms were expected to be more precise and useful to the user.

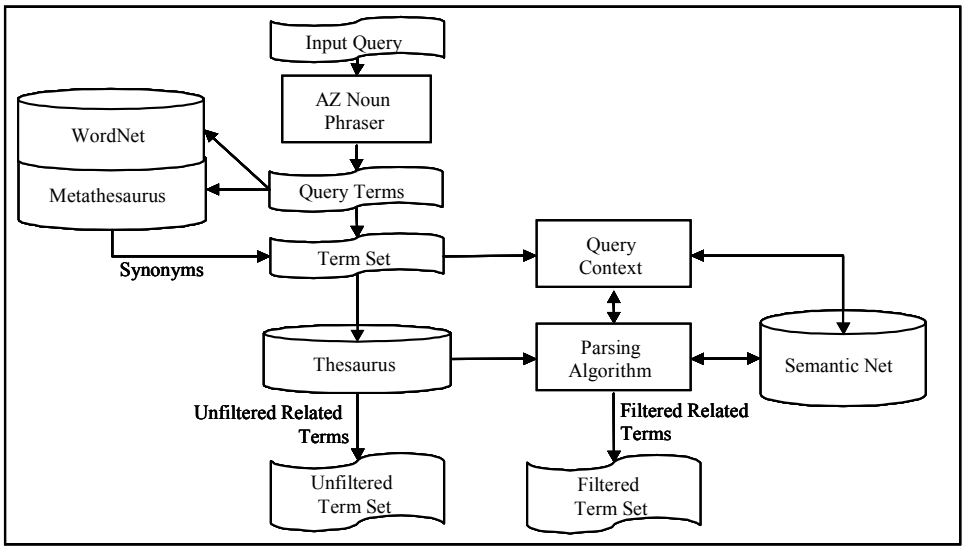


**Figure 2. The Combination of a Medical Thesaurus with Ontologies**

A first evaluation study used 30 real user, cancer related questions, which were automatically submitted to the system. Noun phrases automatically extracted from the questions, synonyms, and thesaurus terms were matched to each query. Two groups of experts, two medical librarians and two cancer researchers, compiled a gold standard of terms for all questions. The two groups compiled substantially different gold standards. The cancer researchers' gold standard contained on average only 6.1 terms. The gold standard of the medical librarians contained on average 17.6 terms, many of which were synonyms and related terms and only the results related to this standard are discussed here. For each question, the automatically retrieved synonyms and related terms were compared against its gold standard. Compared to adding only synonyms, adding additional related terms increased the recall ($p < .05$) but decreased precision ($p < .001$). However, when the parsing algorithm selected terms that fit the query context, recall increased ($p < .05$) without lowering precision. The results indicate that using the UMLS ontology to limit related terms from a probabilistic network improved the precision and relevance of these thesaurus terms (Leroy and Chen, 2001) and that it is possible to automatically construct the semantic context for a natural language sentence. This parsing algorithm is currently being integrated in a medical online search interface (http://http://ai.bpa.arizona.edu/go/medical/MedTextus.html) as a Keyword Suggester. A follow-up evaluation will look at the usefulness of the expansion terms for different types of tasks in a scenario-based user study.

### *Ontologies to Augment Natural Language Processing Techniques*

Natural language processing (NLP) techniques are combined with medical ontologies to automatically and precisely extract textual information from medical texts. Other researchers have used NLP tools ranging from general-purpose parsers to pattern matching for this purpose. However, the complexity of the medical language used instigates parsing errors and problems with overall processing speed that impair full parsers (Park, et al., 2001; Yakushiji, et al., 2001). By using more shallow parsing to extract specific gene, protein and interaction information (Blaschke, et al., 1999; Ng and Wong, 1999; Pustejovsky, et al., 2002), the information can be extracted but is unfortunately limited to a few pre-specified concepts.

A novel parsing algorithm based on English function words, e.g., prepositions and conjunctions, is used to fill generic preposition based templates. By building templates around prepositions, more information is captured, e.g., genes and proteins but also diseases and cell phases. In addition, precision remains high. Figure 3 shows an example for a by-template. A first test with two prepositions (Leroy and Chen, 2002) showed 70% precision and 47% recall. Most approaches to automated extraction of biomedical information report precision between 60% and 90% depending on the different definitions of precision used and also

on the diversity of the extracted information. Recall is limited to sentences containing pre-specified concepts and is therefore very low. Consequently, this new approach has (and the initial test supports this) much better coverage of the relevant concepts and has better scalability.

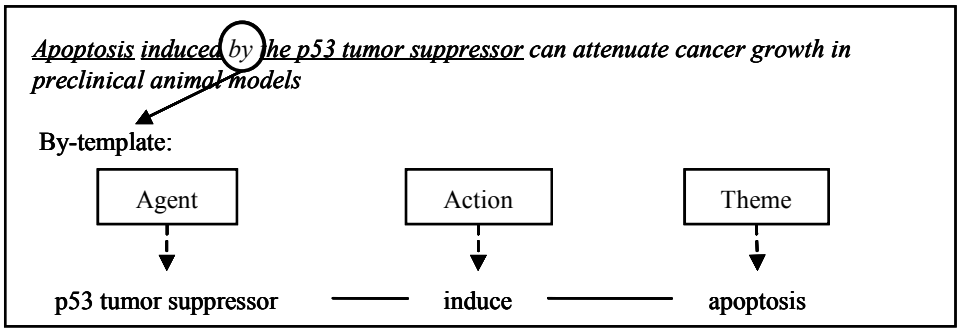


**Figure 3. An Example of the By-Template**

## *GeneScene*

Due to the recent explosion of information in the biomedical field, it is hard for a single researcher to review the complex network involving genes, proteins, and interactions. GeneScene is a toolkit that will assist researchers in reviewing existing literature by representing extracted information in a knowledge map. The process will be completely automated so the extracted information can be kept up-to-date, which is vitally important in this rapidly growing field.

The NLP technique is based on the function word based parser described above. The Arizona Noun Phraser (Tolle and Chen, 2000) has currently replaced the heuristics originally used to extract noun phrases. Since over one third of the errors were due to incorrect noun phrases, precision increased considerably. The UMLS Specialist Lexicon, used for part-of-speech tagging, has replaced WordNet since this lexicon is more appropriate for the medical domain. Finally, cascaded finite state automata are now used to extract the prepositional templates. For conjunctional templates, the UMLS Semantic Net will be applied to establish a semantic context for each sentence to extract only templates with fitting semantic types. Figure 4 provides an overview of the GeneScene framework. Once all knowledge is extracted, a knowledge map will be created and text mining algorithms can be used.

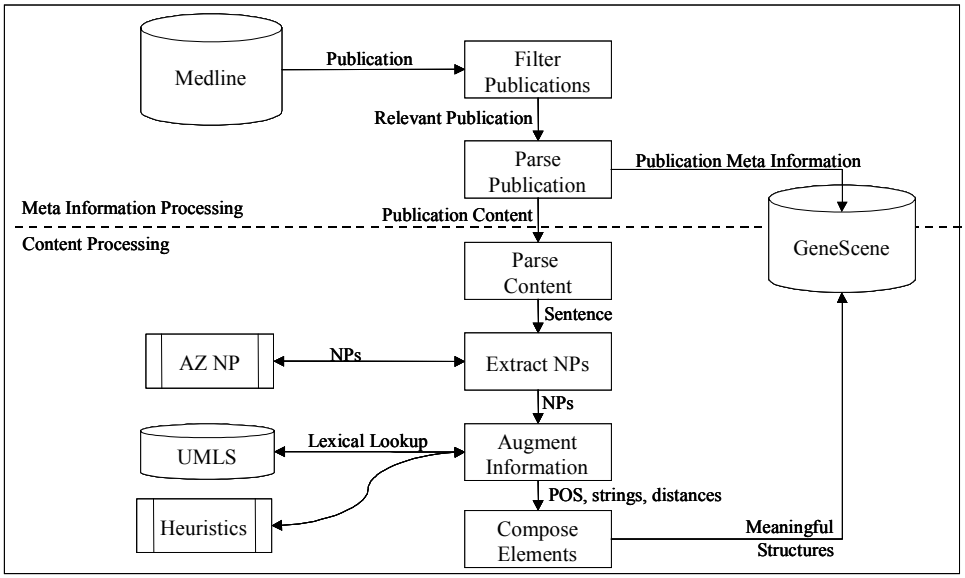


**Figure 4. The GeneScene Framework**

## Conclusion

Three approaches to knowledge discovery were developed and tested in diverse settings. Knowledge discovery can be helped considerable by combining an ontology framework and machine learning algorithms. On-the-fly human judgment is too illusive to guide machine learning. Ontologies are better suited to be combined with machine learning systems. However, the ontology has to be precise and sufficiently relevant to the intended system. Furthermore, it is best to combine the text-based data or information in a bottom-up fashion with the ontology. This helps disambiguate relations between elements, something that troubles top-down approaches. The successful strategies are tested in GeneScene, a complete systems to automate biomedical pathway information extraction from text.

### References


Alavi, M. and Leidner, D.E. "Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues," *MIS Quarterly* (25:1), 2001, pp. 107-136.

Blaschke, C., Andrade, M.A., Ouzounis, C. and Valencia, A. "Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions," *Proceedings of the Seventh International Conference on Intelligent Systems forMolecular Biology*, 1999, pp. 60-7.

Chen, H. "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms," *Journal of the American Society for Information Science* (46:3), 1995, pp. 194-216.

Chen, H. and Lynch, K.J. "Automatic Construction of Networks of Concepts Characterizing Document Databases," *IEEE Transactions on Systems, Man and Cybernetics* (22:5), 1992, pp. 885-902.

Chen, H., Shankaranarayanan, G. and She, L. "A Machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing," *Journal of the American Society for Information Science* (49:8), 1998, pp. 693-705.

Davenport, T.H. and Prusak, L. *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, Massachusetts, 1998.

Fan, W., Gordon, M.D. and Pathak, P. "Personalization of Search Engine Services for Effective Retrieval and Knowledge Management," *Proceedings of the 2000 International Conference on Information Systems (ICIS)*, Brisbane, Australia, 2000, pp. 20-34.

Gruninger, M. and Lee, J. "Ontology: Applications and Design," *Communications of the ACM* (45:2), 2002, pp. 39-41.

Grüninger, M. and Uschold, M. "Ontologies: Principles, Applications and Opportunities (Tutorial)," *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996, pp. 66-137.

Kraft, D.H., Petry, F.E., Buckles, B.P. and Sadasivan, T. "The Use of Genetic Programming to Build Queries for Information Retrieval," *Proceedings of the the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence.*, New York, NY, USA, 1994, pp. 468-473.

Langley, P. and Simon, H.A. "Applications of Machine Learning and Rule Induction," *Communications of the ACM* (38:11), 1995, pp. 55-64.

Leroy, G. and Chen, H. "Meeting Medical Terminology Needs: The Ontology-enhanced Medical Concept Mapper," *IEEE Transactions on Information Technology in Biomedicine* (5:4), 2001, pp. 261-270.

Leroy, G. and Chen, H. "Filling Preposition-based Templates to Capture Information from Medical Abstracts," *Proceedings of the Pacific Symposium on Biocomputing*, Kauai, 2002, pp. 350-361.

Leroy, G., Lally, A.M. and Chen, H. "Implicit User Feedback for Internet Searching," submitted to *Decision Support Systems*.

Ng, S.-K. and Wong, M. "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts," *Genome Informatics* (10), 1999, pp. 104-112.

Park, J.C., Kim, H.S. and Kim, J.J. "Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar," *Proceedings of the Pacific Symposium on Biocomputing*, 2001, pp. 369-407.

Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M. and Cochran, B. "Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations," *Proceedings of the Pacific Symposium on Biocomputing*, 2002, pp. 362-373.

Salton, G. and Buckley, C. "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science* (41:4), 1990, pp. 288-297.

Tolle, K.M. and Chen, H. "Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools," *Journal of the American Society of Information Systems* (51:4), 2000, pp. 352-370.

Yakushiji, A., Tateisi, Y., Miyao, Y. and Tsujii, J. "Event Extraction from Biomedical Papers Using a Full Parser," *Proceedings of the Pacific Symposium on Biocomputing*, 2001, pp. 408-419.