

# INTEGRATING N-GRAM LEXICAL MATCHING WITH A SEMANTIC KNOWLEDGE BASE FOR DOCUMENT ANALYSIS

Barrett R. Bryant \*  
University of Alabama at Birmingham  
bryant@cis.uab.edu

Li Li \*  
Unisys Corporation  
lli@slu.tr.unisys.com

Mark R. Kindl  
Army Research Laboratory  
kindl@airmics.gatech.edu

## Abstract

A combination of existing software tools was investigated for the purpose of analyzing the content of Department of Defense (DoD) documents and classifying them according to official file plans, such as the Modern Army Recordkeeping System (MARKS). The tools used were the Propeller package to perform n-gram lexical matching among documents and the FILAS package to classify documents according to a built-in semantic knowledge base. The approach was tested against a corpus of documents, both paper reports and electronic mail messages, and the results showed that an integration of the two approaches was an improvement over using either of the two individually.

## 1 INTRODUCTION

This work addresses the need to review extensive collections of diverse documents of the United States Department of Defense for the purpose of (1) categorizing them by content and (2) identifying those that can be declassified pursuant to Executive Order 12958. In the Army context, the task of filing a record is for office personnel to determine the appropriate MARKS [MARK93] file number for it according to the file plan. Unfortunately, it is extremely difficult, if not impossible, to classify the records of such significant magnitude and complexity solely by human effort. Automated filing assistance tools, such as the FILAS package [Unde93], which classifies documents according to a built-in semantic knowledge base, are designed for interactive use so the

efficiency is totally unacceptable for intensive application to large quantities (in terms of millions) of documents. Our approach is to integrate 1) the Propeller package [Dama95], a method and software system for gauging similarity among text-based documents and visualizing relationships in data using n-gram lexical matching, and 2) the FILAS package for automated classification of records according to the MARKS file plan. The approach was tested against a corpus of documents, both paper reports and electronic mail messages, and the results showed that an integration of the two approaches was an improvement over using either of the two individually. The findings and methods used should be applicable to other DoD domains as long as a MARKS style file plan can be accommodated for the domain. Furthermore, the system can be improved by further integration of the two software systems.

The organization of this paper is as follows. Sections 2 and 3 introduce the FILAS and Propeller systems, respectively. In Section 4, we discuss how FILAS and Propeller were used in our experiments. We conclude in Section 5.

## 2 FILAS AUTOMATED FILING ASSISTANT

FILAS is an intelligent filing assistant in the File Plan Team system [Unde93]. FILAS understands an organization's file plan and a person's description of records to be categorized or located in the office files. FILAS interacts with a person to aid in assigning the proper filing category to a record. The categories produced by FILAS interactively from a description can be viewed as a tree whose vertices indicate MARKS categories and are ordered by the likelihood of being the correct category. Tests of FILAS with file plans containing 8000 filing categories have shown it to be fast and accurate. However, there are certain limitations of FILAS:

---

\* This research was supported in part by the United States Army Research Laboratory grant number 527199 and carried out at the ARL Software Technology Branch in Atlanta, Georgia.

---

1. FILAS needs the intervention of people in the process of filing a record. To file a document, a user must compose a proper document description in English based on his summarization of the document and input to FILAS. Then the user has to search the category “tree” recommended by FILAS in order to determine a proper category for the document.
2. Some paper records have a subject line that can be used as the description for FILAS. However, most of the subject lines are not accurate enough and the subjects of electronic messages can often be especially misleading to FILAS. Simple extraction of document subject lines would be a naive approach.

The challenge here is to find or develop an efficient method that summarizes a document into a set of keywords based on its content such that they can reduce the search space of FILAS during the filing process.

### 3 PROPELLER N-GRAM LEXICAL MATCHING SYSTEM

The Propeller package [Dama95] consists of several independent tools for text matching, visualization and reduction. The key components are *Acquaintance* and *Parentage*.

Acquaintance is a language-independent tool for gauging topical similarity in unrestricted texts without prior information about document content. The tool uses information derived from  $n$ -grams (consecutive sequences of  $n$  characters) with a vector-space technique that efficiently categorizes a large volume of documents and accommodates context by a well-defined procedure. To use Acquaintance to classify documents, a set of documents, whose categories are known, is used as a set of *references*, against which incoming unknown documents are matched and the similarity between the references and the unknowns are measured by numerical scores.

Parentage is a graphical visualization tool that explains the similarity relations calculated by Acquaintance. Each document is represented as a vertex and the similarity relations are displayed as weighted edges. Among several formal concepts provided by Parentage, *cluster* is most useful for document categorization. A cluster, viewed as a graph, is formed by an unknown document and a set of adjacent reference documents. A cluster of an unknown sample therefore provides evidence of its category. Parentage can also be used to

extract highlight phrases from a document or a cluster of documents. The highlight phrases of an individual document denote the words that set this document apart from the other ones, whereas the highlight words or phrases of a cluster denote the common information among the documents. These highlight phrases are derived from the  $n$ -grams against a background document. By carefully selecting an appropriate background document that contains the common words in the domain, the highlight phrases of a document can be regarded as its subject.

Propeller is essentially a “sublexical” or information theoretic approach to document matching. It is applicable to documents of various topics and formats, such as letters, memos, reports, forms and e-mail. However, Propeller does not handle ambiguities or semantic information and therefore is not sufficient to carry out the categorization task because a document and its MARKS specification are related only at the semantic level rather than by surface lexical patterns.

### 4 USING FILAS AND PROPELLER FOR DOCUMENT ANALYSIS

We studied the categorization of two types of documents: the office paper records, scanned into electronic texts by optical character recognition (OCR), and electronic mail messages, which are henceforth called *unknown* documents. We used Propeller and FILAS in different ways to generate, perhaps through iteration, for an unknown document several sets of category assignments, which can be used as the candidates for its final categorization. We have explored three possible combinations: Propeller only, Propeller followed by FILAS, and a mutual feedback loop between Propeller and FILAS, each described in more detail in the following subsections.

#### 4.1 Propeller Only

In this plan, an unknown document is compared against a set of categorized documents using Acquaintance and the cluster generated by Parentage is sorted to obtain the ranked MARKS categories. The experiment is depicted in Figure 1 and explained below:

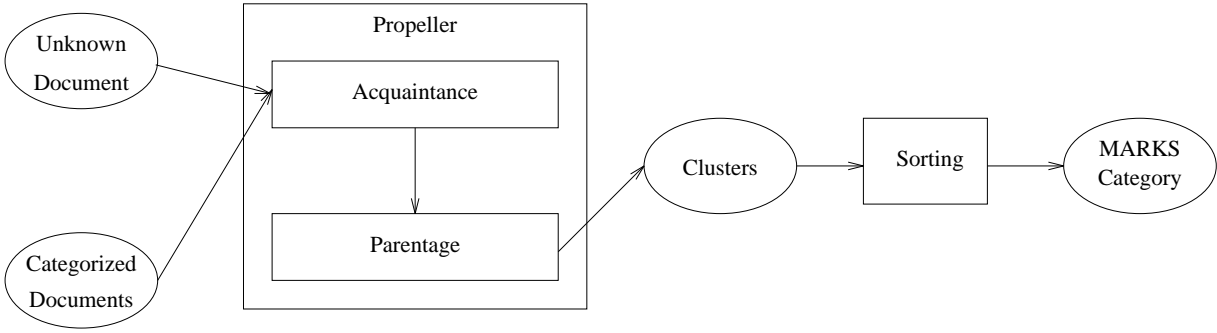


Figure 1: Propeller Only

1. Select  $n$  documents,  $D_1, \dots, D_n$ , of distinct categories where

$$D_i = \{d_j \mid d_j \text{ is a document of MARKS category } M_i\}$$

2. Select a set  $M$  of  $m$  distinct MARKS categories such that  $M_i \in M$  for  $1 \leq i \leq n$ .
3. For each  $d_j$  in  $D_i$ , use  $d_j$  as the unknown document and set  $M$  as the set of reference documents for Acquaintance.
4. Use Parentage to display the cluster of document  $d_j$ .
5. Sort categories in the cluster into descending order based on their rank. Calculate the accuracy of this clustering.

## 4.2 Propeller + FILAS

In this combination, an unknown document is compared against a set of MARKS categories by Acquaintance. The cluster generated by Parentage serves two purposes: (1) provide a set of ranked MARKS categories for the unknown document and (2) derive subject words, which are used by FILAS to deliver the MARKS categories. The experiment is illustrated in Figure 2 and described below:

1. Select  $n$  documents,  $D_1, \dots, D_n$ , of distinct categories where

$$D_i = \{d_j \mid d_j \text{ is a document of MARKS category } M_i\}$$

2. Select a set  $M$  of  $m$  distinct MARKS categories such that  $M_i \in M$  for  $1 \leq i \leq n$ .

3. For each  $d_j \in D_i$ , use  $d_j$  as the unknown document and set  $M$  as the reference documents for Acquaintance.
4. Use Parentage to display the cluster of document  $d_j$ .
5. Sort categories in the cluster into descending order based on their rank. Calculate the accuracy of this clustering.
6. Use Parentage to generate highlight phrases and a filtering algorithm to select subject words from these highlight phrases.
7. Input the subject words to FILAS and record the category tree. Calculate the accuracy of the FILAS category tree.

## 4.3 Propeller + FILAS Iteratively

Propeller and FILAS can also be integrated together into a mutual feedback cycle, as illustrated in Figure 3. The model works as follows:

1. The **Reduction** produces the subject description of the unknown document.
2. FILAS delivers a tree of MARKS categories from the given description.
3. The **Selection** procedure selects reference documents from the known sample documents in light of the FILAS category tree to cut down the number of references.
4. Propeller produces a document cluster and highlights from matching the unknown against the selected references.

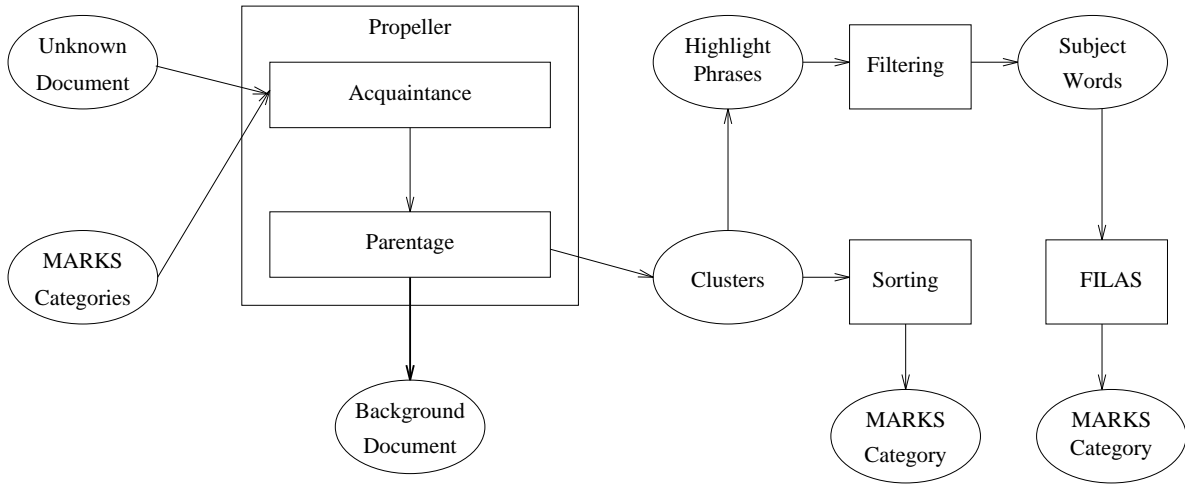


Figure 2: Propeller + FILAS

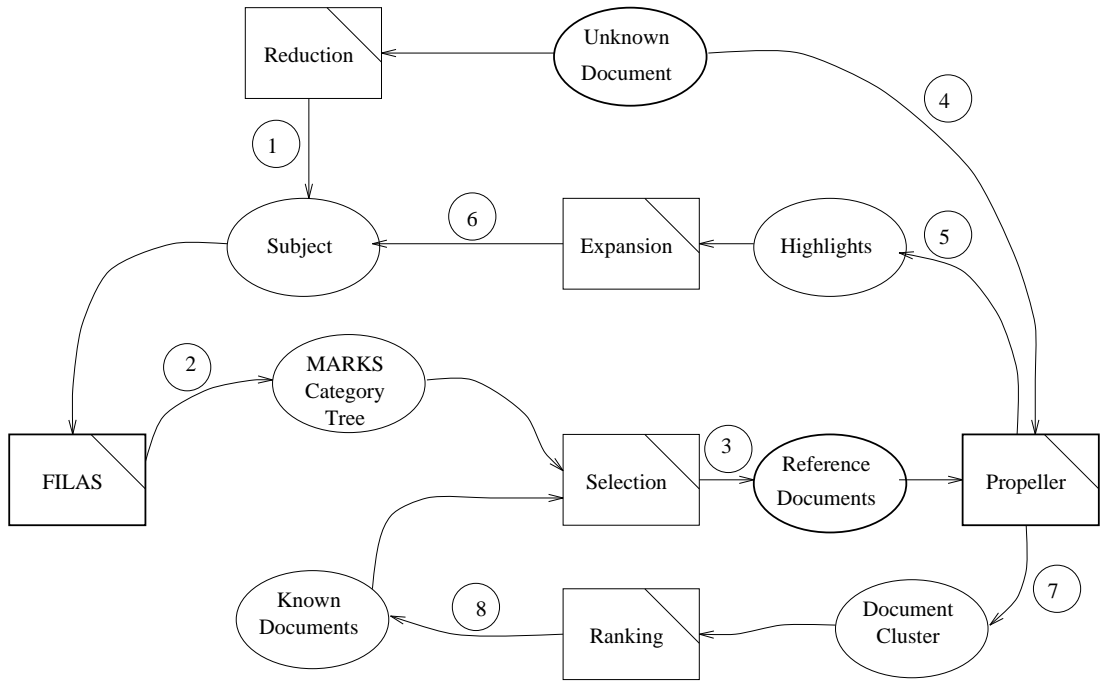


Figure 3: Mutual Feedback Between Propeller and FILAS

5. The documents in the cluster are ranked and used as the known documents for the next round.
6. The highlight phrases are expanded to generate a new subject, which is then fed back to FILAS.
7. The cycle iterates until the **Selection** narrows down the reference documents to an acceptable range. The categories of those references are taken as the candidate MARKS categories for the unknown.

The critical steps of this cycle are **Reduction** and **Expansion**, which are responsible to provide FILAS with accurate subjects. The complete implementation of this is still in progress but prototype studies suggest it has the potential for being more successful than either of the previous two approaches.

## 5 RESULTS/DISCUSSION

We performed experiments using Propeller+FILAS and Propeller Only on large numbers of paper records and e-mail. For simplicity, we use the 1-best test for Parentage clustering: if the correct category is ranked the highest in a Parentage cluster, then the accuracy of this cluster is 1, otherwise it is 0. Our findings are summarized below. Detailed statistics are given in [Brya96].

1. Propeller has moderate performance (74% for paper records and 60% for e-mail) on clustering documents against documents but poor outcome (19% and 7% respectively) on clustering documents against MARKS categories. There are only 1/27 cases (3.7%) in paper records, and 1/28 cases (3.5%) in e-mail, where documents against MARKS outperforms documents against documents. Considering that we used the 1-best test, the performance of documents against documents is satisfactory but certainly needs improvement.
2. Propeller produces better clustering on documents than on e-mail since e-mail in the same category is more ambiguous and less formal than the paper records.
3. FILAS delivers 51.9% hits on documents and 46.4% on e-mail from the subject words derived by the filtering algorithm.
4. If a document can be classified correctly by Propeller (using documents vs. documents), its subject words will more likely lead FILAS to a correct

category than a dead end. On the other hand, if a document is misclassified by Propeller, it is highly possible that its subject words will not work for FILAS either. In order to better constrain FILAS, more sophisticated lexical processing should be employed in deriving subject words from Propeller's highlight phrases.

5. The maximal depth of a FILAS tree is 3, where it either hits a correct category or reaches a dead end. The minimal depth is 0, where FILAS fails to recommend any category from subject words.
6. The average number of categories produced by FILAS for paper records are estimated to be 60% (321.7/536.2) of those produced for e-mail, meaning that the selected subject words will be more accurate (3.67 times more accurate in practice).
7. FILAS produces different rank patterns for paper records and e-mail. Composite N-best tests, whereby we only inspect the first N choices from each FILAS category list, suggest that the subject words of paper records constrain FILAS better than those of e-mail do. However, the different rank patterns of records and e-mail suggest that it might be helpful if we use different N-best values at different levels instead of a uniform one.

For improving the performance, we should inspect the structural properties of various documents, for instance, the transmission relationships of e-mail messages, to derive more evidence for classification.

*Acknowledgements.* The authors would like to thank Marc Damashek and Jim Hoover of the National Security Agency, Bill Underwood of AI Atlanta, Inc., and Dan Hocking of the Army Research Laboratory for their assistance in performing this work.

## References

- [Brya96] Bryant, B. R., Li, L., and Kindl, M., "Integrating N-Gram Lexical Matching with a Semantic Knowledge Base for Document Analysis," Technical Report, Department of Computer and Information Sciences, University of Alabama at Birmingham, May 1996.
- [Dama95] Damashek, M., "Gauging Similarity with n-Grams: Language Independent Categorization of Text," *Science* 267 (February 10, 1995), 843-848.

[MARK93] *The Modern Army Recordkeeping System (MARKS)*, U. S. Army Regulation 25-400-2, Washington, D. C., February 26, 1993.

[Unde93] Underwood, W. E., Laib, S. L., and Cotan, C. W., "MARKS and the File Plan Team," U. S. Army Research Laboratory Technical Report ARL-TR-118, January 1993.