Better Rules, Fewer Features:
A Semantic Approach to Selecting Features from Text.

Catherine Blake and Wanda Pratt
Information & Computer Science
University of California, Irvine
{cblake, pratt}@ics.uci.edu

Abstract

The choice of features used to represent a domain has a profound effect on the quality of the model produced; yet, few researchers have investigated the relationship between the features used to represent text and the quality of the final model. We explored this relationship for medical texts by comparing association rules based on features with three different semantic levels: (1) words (2) manually assigned keywords and (3) automatically selected medical concepts. Our preliminary findings indicate that bi-directional association rules based on concepts or keywords are more plausible and more useful than those based on word features. The concept and keyword representations also required 90% fewer features than the word representation. This drastic dimensionality reduction suggests that this approach is well suited to large textual corpus of medical text, such as parts of the Web.

1. Introduction

Selecting features that are necessary and sufficient is critical if you are to construct a model that can accurately predict future events or describe a problem space. In addition, each individual feature should be **informative;** that is, it clearly captures some aspect of the problem space. Intuitively, models based on informative features will be easier to interpret than models based on uninformative features. For text, the feature representation is tightly coupled to model quality because they are embedded in natural language. Unlike traditional numeric, categorical, and Boolean data types, a textual resource must first be transformed to an alternative representation before a data-mining technique can be applied.

We designed an experiment to understand the relationship between the quality of features used to represent text and the quality of a descriptive model (as measured by plausibility and usefulness). We describe the quality of features in terms of **semantic richness.** For example, *breast cancer* is a disease occurring in a particular part of the body. If a text-mining system represented this phrase using the two individual features *breast* and *cancer*, it would not capture the meaning of the phrase *breast cancer*. Our approach uses a concept feature identified with the entire phrase *breast cancer*, which also captures the semantically equivalent expression *neoplasm of the breast*. We call this approach a semantic approach because it uses a semantic model to determine the features, rather than simply identifying commonly occurring phrases in text. Thus, we say that the concept feature *breast cancer* is semantically richer than the individual features *breast* and *cancer.*

For a model based on text to be valid, our first requirement is that it be **plausible**, that is 'seemingly reasonable or probable' [1]. Clearly, no one will use a system that produces implausible models. For example, neurologists were unwilling to follow decision rules that violated monotonicity constraints, where the neurologist expected an increase (decrease) in an attribute to correspond to an increase (decrease) in a predictive variable[2]. We argue that a model based on text must be plausible before it can be either meaningful or interesting.

Our second requirement is that the generated model be **useful**. By useful, we refer to task-specific usefulness, such as 'would this rule be useful if you were treating a patient with breast cancer?' Although usefulness is a stated goal of data mining [3], it is rarely included as a metric in evaluation, perhaps because it is inherently subjective. Our work explicitly measures the usefulness of a model based on text.

Our hypothesis is that increasing the information content (semantic richness) of features used to represent text will correspond to an increase in the plausibility and usefulness of the descriptive model produced. We used features at three different semantic levels: (1) words (2) manually assigned keywords and (3) automatically selected medical concepts. The model we used was a set of bi-directional association rules. In addition to model quality, our goal was to study the effect of dimensionality when features have varying levels of semantic richness.

2. A Text Mining Scenario

As the amount of text available in electronic form continues to increase at an alarming rate, the tools to

manage these textual resources effectively will become critical. Consider **MEDLINE**, a bibliographical database of medical abstracts from conferences and journals that contains more than 11 million entries[4]. The National Library of Medicine (**NLM**), who maintains this resource, estimates that more than 400,000 additional references will be added during 2001 (8,000 new entries each week). Although access to abstracts or the full-text of articles in more than 4,000 biomedical journals has the potential to be useful for medical researchers, the quantity and unstructured nature of this information often results in information overload. Text mining has the potential to reduce information overload by providing a user with patterns from the underlying text.

Our scenario starts with a medical researcher who wants to learn about breast-cancer treatments. She could search MEDLINE to retrieve bibliographic details of documents related to the topic, however if she spent only 1 minute on each abstract and worked 10 hours a day it would take her a month to read the 19,167 related abstracts. Clearly reading every abstract is infeasible. If her task related to a specific treatment, then she would provide additional constraints to narrow the search space. However, her goal is to learn about treatments; thus, it is unreasonable to assume that she knows the names of the available treatments.

A text-mining system would enable her to understand treatments by providing a model of the important relationships within the published scientific literature. The model, which would be at a finer level of granularity than the entire abstract, might identify a co-occurrence between a treatment and its side effects. For example, *Docetaxel*, a chemotherapy drug that destroys cancerous cells, sometimes destroys cells that grow at a fast rate, such as those responsible for hair growth; thus, a patient may suffer from Alopecia (hair loss). A rule associating *Docetaxel* with *alopecia* would be useful to our medical researcher because it would help her understand the nature of the treatments available for breast cancer. Further, co-occurrences previously unreported in any individual study would be of particular interest. Let us use an example from a different domain. Tengs and Osgood found evidence that impotence correlated with smoking. They discovered this correlation by using clinical trials that did not specifically analyze the relationship between impotence and smoking, but rather studied impotence, and happened to report tobacco usage[5]. Although they used a manual rather than an automated approach, we believe that a text-mining system should also identify a correlation between concepts that are not the primary focus of an individual study.

3. Related Work

Identifying informative features from natural language (text) can be difficult; thus, existing approaches use semantically poor features, such as words[6-14]. This approach has the advantage of being domain independent and easy to implement; it has the disadvantage of producing the same number of attributes as the size of the vocabulary. The Apriori algorithm requires potentially 2^m item sets where m is the number of terms in the vocabulary (see section 5). Although it is unlikely that every term will appear in every textual resource, (the condition that makes 2^m item sets necessary in practice), the number of features can seriously impede the application of this data-mining algorithm.

A representation that does not account for natural language characteristics, such as synonymy or polysemy, could cause a data-mining system to generate a misleading model. Consider the phrase *hair loss*, which is synonymous with the medical term *alopecia*. The actual number of times that this concept occurs is the number of times that either *hair loss* or *alopecia* occurs in the text, but a word-based representation would distribute the count between the three features *hair*, *loss,* and *alopecia.* Thus, the word-based count would be smaller than the actual number of occurrences of the medical concept *alopecia*. The word-based representation could also over estimate the count for a concept. The word-based feature count for *hair* would also include the number of times that the expression *hair loss*, *hair gain* and *hair color* occurred in the collection. In contrast, a concept representation would unify the expression *hair loss* with *alopecia* and thus account for synonymy. Although you could augment a word-based system with list of synonymous word pairs such as *nausea* and *queasiness,* it is unclear if this would be effective in removing the influence of synonymy that is present in text.

Other researchers have explored the use of manually assigned keywords. For example, Feldman and colleagues used keywords as features for the generation of association rules[15-17]. However, they did not evaluate the effect of their feature-selection method on the plausibility or usefulness of the generated rules, but rather they determined the time it took to generate the rules. The drawbacks of approaches that use manually assigned keywords are that (1) it is time consuming to manually assign the keywords (2) the keywords are fixed (i.e., they do not change over time or vary based on a particular user) (3) if the keywords are manually assigned, they are subject to discrepancy (4) the textual resources are constrained to only those that have keywords.

Other researchers have used representations of medical concepts that are similar to ours for automatically determining a diagnosis category based on a textual description of a patient's clinical report[18]. They used a physician's diagnosis, also based on the clinical report, as a gold standard, and found that the automatically determined diagnoses based on concepts were more accurate than those based on words. We use the same extensive medical knowledge base; however, our

approach identifies concepts from free-form text in scientific abstracts, rather than from clinical reports. Their work also relates more closely to a text categorization task than the text-mining scenario outlined in section 2.

Other studies have examined the effect of feature selection from text on learned models[19,20]. However, these studies use statistical techniques with no semantic component for feature-selection. In addition, the studies measured the performance of their approach on a document-categorization task, rather than the text-mining task in section 2.

Other related research has focused on constructing techniques to improve the quality of text-mined association rules. Most of these approaches first generate a set of rules, and then apply pruning or ranking techniques, such as 'interestingness' [14,21-25]. Unlike these approaches, which consider each rule individually, we focus on improving the interestingness for the set of rules. That is, on average, rules based on semantically richer representations would be higher than rules based on a semantically poorer representation. We consider plausibility to be a necessary but not sufficient condition for interestingness. Several systems enable users to provide expectations to a system, and then rank rules with respect to how they differ from the users' expectations[24,25]. Although this approach enables a rule ranking for a particular user, it requires considerable effort from users to specify their expectations, such as *the overall instruction ratings are higher than the overall course ratings*[22]. Users are also required to give the degree of belief, a probability; however, there is strong evidence that people are not good at estimating such probabilities [26].

Although we agree that ordering the rules could be valuable, we advocate ranking in conjunction with using informative features. We would be surprised if a plausible and useful rule would ever use the feature 'is'; yet, according to the results of one study based on the χ^2 ranking, the correlation between the features *government*, *is*, and *number* was ranked the most interesting [27].

4. Feature Extraction

In our experiments, we varied the semantic richness of the features used to represent the title and abstract of each bibliographic reference. We now describe the process that we used to generate each of the feature sets.

4.1. Concept Features

Few researchers would claim that a word representation is optimal, but the difficulty of automated natural-language understanding has limited our ability to use a richer representation scheme. Unlike many natural-language systems, our system does not use a part-of-speech tagger to identify candidate phrases. Instead, it uses a heuristic approach to break each sentence in the document's title and abstract into a set of clauses. The system uses stopwords; words, such as *the*, *in* and *it*, that occur frequently in text but are not meaningful, to separate the sentence into clauses. We used a generic list of stopwords that was developed for information retrieval purposes[28], and we added to this list numbers, days of the week, and month names. We also developed and included a second set of 31 medical stopwords that occur frequently but are not meaningful, such as *study* and *test*.

We used an existing knowledge base, the **Unified Medical Language System (UMLS)** to map each clause to a medical concept. The UMLS (which was created and is maintained by the NLM) consists of three components: (i) a semantic network that links each concept to one of 132 high-level concepts called semantic types (e.g. the concept Tamoxifen, a chemotherapy drug, has a semantic type of Organic Chemical), (ii) the **Metathesaurus**, a medical thesaurus that contains synonymy and hierarchical links among about 800,000 concepts and 1.9 million concept names and (iii) the SPECIALIST lexicon, which provides lexical information on 140,000 concepts. For example, the UMLS maps the clause *MR-guidance* to the concept *magnetic resonance imaging guidance*.

The last step was to impose semantic constraints on the concepts provided by the UMLS so that the concepts related to findings and treatments. Based on the semantic type hierarchy, we, selected the following types as suitable: *Therapeutic or Preventive Procedure*; *Sign or Symptom*; *Pharmacologic Substance*; *Neoplastic Process*; *Amino Acid, Peptide, or Protein*; *Antibiotic or Organic Chemical*.

```
1. ConceptList ← θ
2. For each sentence in the title and abstract
3.   For each clause in the sentence
4.     concept List ← + UMLS concept using the clause
5.   End for
6. End for
7. ConceptList ← constrain using semantic types
```

Figure 1 – Concept Extraction Process

Any knowledge-based approach such as ours is clearly dependent on the comprehensiveness and quality of the knowledge it is based on. To assess the quality of our clause-concept mappings, we manually reviewing the original 367 concepts provided by the UMLS. We identified 18 (5%) blatant mismatches and have since notified the NLM. We also noticed that some clauses were mapped to a concept that was more specific or more general than the original clause (e.g., *tolerated* was mapped to the concept *maximum tolerated dose*, and *hematologic recovery* to the concept *recovery from disease*). Although such errors affect model quality, we did not remove these concepts from our feature set.

4.2. Keyword Features

Employees of the NLM assign 10-12 keywords from a controlled vocabulary to each bibliographic reference in the MEDLINE database. Documents are indexed, organized and retrieved using this medical ontology (the **Medical Subject Headings**, or **MeSH**). In addition to the 19,270 MeSH terms, the employees also use 3-4 of the 800 available subheadings (or qualifiers) to index the medical literature. A qualifier provides specific details about the application of a MeSH term in a document. For example, the qualifier *drug effects*, when added to the MeSH term *Liver*, indicates that the article or book is not about the liver in general, but rather is specifically about the effect of drugs on the liver. We used both the MeSH terms and qualifiers as our keyword features.

4.3. Word Features

Researchers generally use word features to represent text [9,11,12,27]. We used the same stopwords as those used to generate concepts, that is a generic set of stopwords [29] augmented with numbers, months, days of the week and 31 medical stopwords.

The approach most often used is to remove stopwords, and then do word **stemming**, a process that removes a word's prefixes and suffixes. Instead of using a generic stemming algorithm, such as Porter's, we used the Lexicon Variant Generator (lvg), a stemming tool provided by NLM that was specifically designed for the medical domain. In addition to removing suffixes and prefixes (such as unifying both *analyzed* and *analyzing* to *analyze*), lvg unifies morphological changes (such as transforming *wound* to *wind*). We applied pre-processing operations in the following order: convert to lower-case, remove stopwords, strip genitive or possessive, strip punctuation, uninflect, canonicalize. We then removed duplicates of the same canonical form from each abstract.

5. Generation of Association Rules

Following the scenario in section 2, we started with the 19,167 abstracts from MEDLINE that relate to the query *breast cancer treatment*. We then selected the most recent 100 abstracts from this set. We discarded 9 abstracts because they did not have MeSH terms that corresponded to a treatment, and we did not want irrelevant documents to bias our results. The title and abstract of the remaining 91 articles were used to generate the word and concept features. We used the MeSH terms and qualifiers of those documents as the keyword features. Table 1 shows the rules that we used in our experiment.

	Rules based on Word Features	Rules based on Keyword Features	Rules based on Concept Features
1.	Axillary ↔ Background	Drug therapy ↔ Human	Doxorubicin ↔ Prekallikrein
2.	Metastatic ↔ Toxicity	Disease-free survival ↔ Survival analysis	Alopecia ↔ Methotrexate
3.	Grade ↔ Response	Antineoplastic agents, phytogenic ↔ Paclitaxel	Chemotherapy-Oncologic Procedure ↔ stage IV breast cancer
4.	Chemotherapy ↔ Result	Drug therapy ↔Therapeutic use	Carcinoma of Breast ↔ Vinorelbine
5.	Advance ↔ Phase	Lymph Nodes ↔ Radionuclide Imaging	Anorexia ↔ Progressive disease
6.	Blue ↔ Detection	Cyclophosphamide ↔ Methotrexate	Nausea ↔ Stable disease
7.	Cancer ↔ Trial	Antineoplastic agents, phytogenic ↔ Infusions, intravenous	Docetaxel ↔ Neoplasm Metastasis
8.	Milligram ↔ Toxicity	Antineoplastic agents, combined ↔ Breast Neoplasms	Cyclophosphamide ↔ Lymphocyte antigen CD69
9.	Conclusion ↔ Studied	Support, Non-U.S. Govt ↔ Treatment Outcome	Fatigue ↔ Granulocyte Colony-Stimulating Factor
10.	Conclusion ↔ Trial	Adolescence ↔ Lung Neoplasms	Docetaxel ↔ Pain
11.	Common ↔ Tamoxifen	Aged ↔ Drug therapy	Fatigue ↔ Nausea
12.	Advance ↔ Grade	Etiology ↔ Radiotherapy	Doxorubicin ↔ Paclitaxel
13.	Conclusion ↔ Dose	Adolescence ↔ Nausea	Enterotoxin F, Staphylococcal ↔Stable disease
14.	Median ↔ Respectively	Postmenopause ↔ Tamoxifen	Paclitaxel ↔ Vinorelbine
15.	Nausea ↔ Vomit	Lung neoplasms ↔ Nausea	Nausea ↔ Stage IV Breast Cancer
16.	Cancer ↔ Method	Antineoplastic Agents, Combined ↔ Drug Therapy	Axillary Lymph Node Dissection ↔ Secondary Malignant Neoplasm of Lymph Node
17.	Baseline ↔ Questionnaire	Disease-free survival↔Prospective studies	Gemcitabine ↔Vomiting
18.	Breast ↔ Cancer	Aged, 80 and over ↔ Postmenopause	Lymphocyte antigen CD69 ↔ Myalgia
19.	Dose ↔ Phase	Aged ↔ Therapeutic use	Enterotoxin F, Staphylococcal ↔ Gemcitabine
20.	Progressive ↔ Stable	Female ↔ Pathology	Anthracycline Antibiotics ↔ Epirubicin

Table 1 The bi-directional association rules used in our evaluation.

To identify patterns in the text, we used the popular data-mining technique of generating association rules. An **association rule** is of the form A→B, which means "thepresence of A implies the presence of B", where A is the set of antecedents and B is the consequent set. An individual rule identifies a co-occurrence between the antecedent and consequent sets. Each association rule has an associated level of support and confidence. The **support** is the probability that both A and B occur in a textual resource. For example, if 89% of the abstracts contained both the words *breast* and *cancer*; then the support of the association rule *breast→cancer* is 89%. The **confidence** is the probability that B will occur given that A has already occurred. We constrain the rules to those with a single feature in each of the antecedent and consequent sets. We define a **bi-directional association rule** (indicated by ↔) as one that satisfies the support and confidence levels in both directions (i.e., both A→B and B→A have support and confidence greater than the minimum that was set by the user).

We used Borgelt's implementation of the Apriori algorithm to generate association rules on each feature set separately[30,31]. The computationally intensive component of the Apriori algorithm is to identify the **item sets,** those features that occur with a frequency greater than the specified minimum level of support. The algorithm then generates rules that satisfy the minimum level of confidence based on these item sets. We used only bi-directional rules in our experiments.

6. Experiments

Our goal was to determine whether using features of differing semantic richness has an effect on the plausibility or usefulness of bi-directional association rules based on those features. We used two perspectives on plausibility and usefulness, a physician's and the consumer information available from the American Cancer Society's website on breast cancer treatment [32].

	Support (%)	Confidence (%)	Bi-directional Rules
Word	6	48	213
Keyword	4	32	104
Concept	2	16	156

Table 2 – We lowered the minimum support and confidence because the semantically richer representations had fewer features.

Fewer features are required to represent an abstract as a set of concepts, rather than as a set of words. To produce the same number of rules, you would have to lower the support so that approximately the same number of abstracts would satisfy the minimum support and confidence constraints. Therefore, to select a subset of rules for our experiment, we imposed different minimum support and confidence levels for each feature set. We did not select the rules with the highest support and confidence because there is growing evidence that these metrics do not capture interestingness [14,21,22]. We did attempt to control for support and confidence by generating a set of approximately 100 to 200 bi-directional association rules for each feature set. We then randomly selected 20 rules from each feature set for inclusion in the evaluation (see Table 1 for the rules used in this evaluation). We considered three ways of controlling for support and confidence: (1) fix support and lower confidence (2) fix confidence and lower support (3) fix the ratio of support and confidence. As we did not have a principled way to determine the relative importance of support and confidence, we chose to fix the ratio between the two parameters. We set the ratio based on the default values in Borgelt's implementation [30] to 1:8 (see Table 2 for specific values).

6.1. Assessment

We asked an experienced physician to evaluate the bi-directional association rules in Table 1. We provided the physician with definitions of the medical terms used in the associations because he was not an expert in breast cancer. We assumed that if there was a plausible relationship between the features, then a physician could write down that relationship. Thus, we asked him to state any relationship that he found. We instructed him to answer question B based on plausibility and question C, based on usefulness using a scale of *not at all*, *not really*, *neutral*, *mostly,* and *definitely*. We used the same scale for question D, which attempted to uncover the novelty of the rule. Figure 2 shows the exact questions asked.

(1) Physician Questions:

(A) The relationship between these concepts could be …
(B) Do you agree that the relationship in (A) is plausible?
(C) How useful would knowing this relationship be if you were treating a patient with Breast Cancer?
(D) Do you agree that this correlation contributes to scientific knowledge on Breast Cancer treatment?

(2) Consumer Questions:

(A) The relationship between these concepts could be …
(B) Do you agree that the relationship in (A) is strongly implied?
(C) How useful is the relationship in (A) with respect to Breast Cancer Treatment?

Figure 2 – Questions asked for each bi-directional association rule to measure plausibility and usefulness from a physician's perspective (1) and with respect to consumer information on available breast cancer treatment (2).

Our second perspective on the plausibility and usefulness of the rules was with respect to basic health information.

We used the American Cancer Society's (ACS) web page of breast cancer treatments as our gold standard. For each rule, we searched the ACS web page for any information that would relate the antecedent and consequent features. As with the first experiment, responses to (B) and (C) were expressed using a scale of *not at all*, *not really*, *neutral*, *mostly,* and *definitely*. Figure 2 shows the exact questions that we asked. The person doing this evaluation (the first author) was not familiar with breast cancer treatments, so they also referred to the definitions provided to the physician.

7. Results and Discussion

Our preliminary results indicate that bi-directional association rules based on keyword or concept features are more plausible and more useful than association rules generated on words. Although this increase is not statistically significant, our concept and keyword representation had the additional benefit of 90% fewer features than a word representation.

We start with an example of two plausible useful rules generated with concept features: *alopecia↔Methotrexate* and *alopecia↔Tamoxifen*. Both Methotrexate and Tamoxifen are chemotherapy drugs associated with hair loss. Neither the word nor the keyword representations identified this relationship with our minimum support and confidence (see Table 1) even though the word features *hair*, *loss* and *alopecia* occurred in 2, 7 and 6 abstracts respectively. In contrast, the concept representation detected that *hair loss* is synonymous with *alopecia* and accurately recorded that this concept occurred 8 times in our set of abstracts; thus, the concept approach successfully identified this plausible, useful rule. We recognize that eventually *hair↔Methotrexate* would have emerged from the word-based rules if we lowered the minimum support and confidence. However, it is unlikely that a user would be willing to sift through many low quality rules to locate this relationship.

7.1. Physician Assessment

We grouped the physician's plausibility ratings of *mostly* and *definitely* into the general category of *plausible*. Correspondingly, we considered rules to be implausible if he rated them *not at all*, *not really* or *neutral*. We used the same approach to group usefulness ratings.

The results show that rules based on concept features produced more useful rules than those based on either keyword or word features. Conversely, he considered rules based on keyword features to be more plausible than rules based on concept or word features (see Table 3). We use an example to illustrate that it may be reasonable for the physician rate a rule as useful, even though he did not consider the rule plausible. The physician rated the rule between *blue* and *detection*, which was generated from word-based features, as *not really plausible*, but *definitely useful*. His possible relationship was 'Color of test marker and diagnosis of disease' and he annotated 'if it existed' next to the usefulness question. Although plausibility estimates the degree to which the current medical literature supports the association, usefulness reflects the value of a rule if it did exist. In this example, the physician doubted the plausibility of the relationship but recognized its usefulness if it existed.

	Plausible (%)	Useful (%)	Plausible and Useful (%)
Word	10 (50)	6 (30)	5 (25)
Keyword	17 (85)	11(55)	10 (50)
Concept	13 (65)	13 (65)	9 (45)

Table 3 – Plausible and useful ratings of bi-directional association rules with respect to a physician.

The physician was unable to specify any relationship for 10 of the 60 associations. Of the ten, seven were from word features, thus suggesting that word-based associations are neither understandable nor plausible. Although we used lexical tools that were specifically developed for the medical domain, only 5 of the 20 associations rules for word features were considered to be both plausible and useful by the physician.

To express the quality of the associations, we used the information retrieval metric **precision**, which in this case is the number of association rules that are plausible and useful divided by the total number of rules evaluated. The precision of rules with respect to both plausibility and usefulness was 25%, 50% and 45% for word, keyword, and concept features respectively. Thus, the concept and keyword features appear to improve the quality of generated associations.

In addition to evaluating the plausibility, we attempted to capture how novel or insightful each association was, by asking the question 'Do you agree that this correlation contributes to scientific knowledge on breast cancer treatment?' If we group *mostly* and *definitely* into the insightful category, the physician rated 5 rules as insightful that were generated on word features and 8 rules each for keyword and concept features. After inspecting these rules, however, we believe that the physician misinterpreted this question. For example, the physician rated the association between *Docetaxel* (a chemotherapy drug) and *Neoplasm Metastasis* (when cancer spreads to other parts of the body) as a contribution to scientific knowledge, but the American Cancer Society's web page lists *Docetaxel* as a known chemotherapy drug for treating breast cancer; thus, we believe this rule is not novel. This discrepancy reflects the difficulty in developing accurate methods to measure subjective attributes such as novelty or insightfulness.

7.2. Consumer Health Assessment

When compared to consumer information, rules generated from concept features were more plausible and more useful than associations generated using either word or keyword features. The plausibility and usefulness of associations with respect to basic information on breast cancer treatment was overall slightly lower than that from a physician's perspective. Precision with respect to usability and plausibility was 15%, 30% and 85% for word, keyword and concept features respectively. When considering only plausibility, precision increases to 40%, 60% and 90%. The precision when considering usefulness only, is close to the precision of both metrics, specifically 15%, 35% and 85% for word, keyword and concept representation respectively.

	Plausible (%)	Useful (%)
Word	8 (40)	3 (15)
Keyword	12 (60)	7 (35)
Concept	18 (90)	17 (85)

Table 4 – Plausible and useful ratings of rules with respect to consumer health information.

No plausible explanation could be found between the antecedent and consequent for twenty-two of the sixty bi-directional association rules. Twelve of these associations were based on word features. Eight associations were based on keywords and only the remaining two were based on concept features. We assume that there are more unexplained rules in this survey than for a physician because of a lack of formal medical training by the person doing the assessment. We used lexical tools specifically developed for the medical domain, but only 3 of the 20 associations rules for word features were considered useful.

7.3. Dimensionality Reduction

Although the increase in the number of plausible and useful rules is encouraging, the results are not statistically significant. The improvement is more impressive when you consider the drastic reduction in the number of features required to represent the problem space. We summarize the dimensionality reduction in Table 5.

We found that a keyword representation reduced the average number of features by 26% from 76 to 20. The concept representation reduced the average number of features required to represent the title and abstract by 90% to only 8. Both the concept and keyword representations required 84% fewer distinct terms. This reduction has important implications to the running time of computationally expensive data-mining techniques. Data-mining algorithms often require data in a matrix format. Storing the 91 titles and abstracts in this experiment, as a matrix would require 194,012 cells if you used word features and 31,759 cells if you used concepts. This reduction becomes increasing important if all 19,167 articles related to breast cancer treatments were used, or if the full-text of the articles was considered.

	Average unique features	Distinct terms	Abstract-Feature Pairs
Word	76	2132	6932
Keyword	20	351	1856
Concept	8	349	1161

Table 5 – Feature selection has a drastic effect on the dimensionality required to represent the domain.

The concept representation reduced the number of abstract-feature pairs by 83% compared to the keyword approach that had a 73% reduction. Despite smaller space requirements, rules produced using keyword and concept features were more plausible and useful than rules generated over words.

8. Future Work

As we mentioned in the related work section, several researchers have explored approaches to rank rules based on interestingness. We also plan to extend our approach to prune or rank rules based on semantic information, such as the semantic relations from the UMLS. Association rules show only that a correlation has occurred between concepts. We are planning to augment this relationship with semantic information. Consider rule 17 in Table 1, where *Gemcitabine* has a semantic type of *Pharmacologic Substance* and *Vomiting* has the semantic type *Sign or Symptom*. A valid semantic relationship between these two semantic types is *treats*: a *Pharmacologic Substance* ***treats*** a *Sign or Symptom*. We will continue to explore the subjective qualities of a model by interviewing additional physicians. Finally, we plan to conduct experiments on the full text of a document instead of using only the title and abstract. Such experiments will improve our understanding of the impact of the dimensionality reduction with a concept representation.

9. Conclusion

Our hypothesis was that increasing the semantic richness of features used to represent text would have a positive effect on the plausibility and usefulness of a set of bi-directional association rules. Our initial findings support this hypothesis. Specifically, our physician found only 25% of the rules based on word features to be useful and plausible compared with 50% and 45% for keywords and concepts respectively. It was unclear which of the two semantically richer representations were preferred however, because the physician evaluated rules based on keywords to be the most plausible, and rules based on

concepts to be the most useful. The consumer information analysis also supported our hypothesis; the concept-based rules were clearly more plausible and useful than either the keyword or the word representations.

Although the increased model quality was not statistically significant, the 90% reduction in the number of features suggests that the semantically rich keywords or concepts features will enable association rules to be generated more efficiently. The keyword representation constrains the suitable text to those with keywords, but this constraint does not apply to the concept representation. Thus, the concept approach would be suitable to apply to any large corpus of medical text, such as portions of the web.

10. Acknowledgements

We thank Dr. Tony Greenberg, Craig Evans, and Henry Wasserman for their help with the evaluation. This work was supported by the University of California's Life Science Informatics grant #L98-05.

11. References

[1] H. W. Fowler and F. G. Fowler, *The Concise Oxford Dictionary of Current English*, 9th ed. Oxford: Clarendon Press, 1995.

[2] M. Pazzani, S. Mani, and W. R. Shankle, "Comprehensible knowledge-discovery in databases", 19th Annual Conference of the Cognitive Science Society, pp.235-238, 1997.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press, 1996.

[4] National Library of Medicine, Available at http://www.nlm.nih.gov, 2001.

[5] T. Tengs and N. D. Osgood, "The link between smoking and Impotence: Two Decades of Evidence", *Preventive Medicine*, vol. 32, pp.447-452, 2001.

[6] R. Feldman, "Practical Text Mining", PKDD-98, p487, 1998.

[7] R. Ghani, R. Jones, D. Mladenic, K. Nigam, and S. Slattery, "Data Mining on Symbolic Knowledge Extracted from the Web", *KDD-2000 Workshop on Text Mining*, 2000.

[8] U. Hahn and K. Schnattinger, "Knowledge Mining from Textual Sources", CIKM`97, Las Vegas, pp.83-90, 1997.

[9] B. Lent, R. Agrawal, and R. Srikant, "Discovering Trends in Text Databases", KDD'97, pp.227-230, 1997.

[10] U. Y. Nahm, "Text Mining with Information Extraction: Mining Prediction Rules from Unstructured Text", Thesis proposal, University of Texas, Austin, 2001.

[11] G. W. Paynter, I. H. Witten, S. J. Cunningham, and G. Buchanan, "Scalable browsing for large collections: a case study", 5th Conf. Digital Libraries, Texas, pp.215-218, 2000.

[12] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan, "Text mining: a new frontier for lossless compression", Data Compression Conference, pp.198-207, 1999.

[13] H. Ahonen, O. Heinonen, M. Klemettinen, and A. I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections", IEEE Form of Research and Technology Advances on Digital Libraries, pp.2-11, 1998.

[14] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo., "Finding Interesting rules from Large Sets of discovered association rules", CIKM'94, Maryland, USA, pp.401-407, 1994.

[15] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)", ECML-95 Workshop on Knowledge Discovery, Crete, Greece, pp.175-180, 1995.

[16] R. Feldman, I. Dagan, and H. Hirsh, "Mining text using keyword distributions", *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, vol. 10, pp.281-300, 1998.

[17] R. Feldman and H. Hirsh, "Mining associations in text in the presence of background knowledge", KDD-96, Portland, USA, pp.343-346, 1996.

[18] A. Wilcox, G. Hripcsak, and C. Friedman, "Using Knowledge Sources to Improve Classification of Medical Text Reports", (poster) *KDD-2000 Workshop on Text Mining*, 2000.

[19] J. L. Goldberg, "CDM: an approach to learning in text categorization", *International Journal on Artificial Intelligence Tools (Architectures, Languages, Algorithms)*, vol. 5, pp.229-153, 1996.

[20] Y. Yang and J. P. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", ICML'97, 1997.

[21] A. Silberschatz and A. Tuzhilin, "On Subjective Measures of Interestingness in Knowledge Discovery", KDD'95, 1995.

[22] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems", *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, pp.970-974, 1996.

[23] R. J. Bayardo and R. Agrawal, "Mining the Most Interesting Rules", KDD'99, San Diego, pp.145-154, 1999.

[24] B. Padmanabhan and A. Tuzhilin, "Unexpectedness as a Measure of Interestingness in Knowledge Discovery", *Decision Support Systems*, vol. 27, pp.303-318, 1999.

[25] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations", KDD'99, CA, pp.125-134, 1999.

[26] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases", *Science*, vol. 185, pp.1124-1131, 1974.

[27] S. Brin, R. Motwani, and C. Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Dependence Rules", KDD'98, pp.39-68, 1998.

[28] M. Sanderson, "Stop word list", Available at http://www.dcs.gla.ac.uk/idom/ir_resources/, 1999.

[29] S. J. Nelson, W. D. Johnston, and B. L. Humphreys, "Chapter 11: Relationships in Medical Subject Headings (MeSH)", Available at http://www.nlm.nih.gov/mesh/meshrels.html, 2001.

[30] C. Borgelt, "Apriori Implementation", Available at http://fuzzy.cs.uni-magdeburg/~borgel, 1999. .

[31] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and I. Verkamo, "Fast Discovery of Association Rules," in *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, 1995.

[32] American Cancer Society, "Treatment of Breast Cancer Consumer Information", Available at http://www3.cancer.org/cancerinfo, 2001.