Appearing in Lecture Notes in Computer Science. Published by Springer-Verlag

Using Background Contextual Knowledge For Documents Representation

Arkadi Kosmynin and Ian Davidson

CSIRO Division of Information Technology, 723 Swanston Street, Carlton Victoria, Australia 3053

[arkadi.kosmynin | ian.davidson]@mel.dit.csiro.au

Abstract. We describe our approach to document representation that captures contextual dependencies between terms in a corpus and makes use of these dependencies to represent documents. We have tried our representation scheme for automatic document categorisation on the Reuters' test set of documents. We achieve a precision recall break even point of 84% which is comparable to the best known published results. Our approach acts as a feature selection technique that is an alternative to applying the techniques from machine learning and numerical taxonomy.

1. Introduction and Motivation

How to represent documents is central to the problems of information retrieval, document categorisation and information filtering (which we will collectively refer to as document processing problems). There have been many approaches to generate document representations using techniques from statistics, knowledge based systems and natural language processing. Our approach is from the first area.

It is fairly clear that document representation is important, however knowing what information to capture in the representation is a difficult question. Winston from the A.I. community states that problem representation is the most important aspect required to solve a problem [1]. The representation should be conducive to how the problem will be solved. This is often difficult to achieve as a mechanism to understand how to solve a problem is not often available. However with document processing problems we can observe how a human solves these problems and try to imitate this on a computer.

It would appear from human problem solving studies [2] that prior or background knowledge of the problem is essential. Humans store patterns of dependencies and then rely on these dependencies to simplify and solve the problem [2]. This inherently emphasises the salient details and removes non value adding details of the problem. We believe the approach of making use of background information not explicitly contained in the document is useful for solving document processing problems. A researcher interested in optimisation techniques scanning a scientific journal knows that the title of papers containing the terms, “simulated annealing”, “genetic algorithms” and “hill climbing” (all optimisation techniques) are relevant to his search. The researcher uses prior knowledge on the subject to solve the problem. Based on this notion we have attempted to develop a document representation technique that can capture background knowledge for a corpus and then use this knowledge to generate more useful documents representations.

The representation of documents has been investigated by Lewis [3] and Salton [4]. The typical document representation techniques use binary predicates to indicate the presence of (sometimes stemmed) words occurring in the document. Each binary predicate is a feature of the document. The collection of all unique words in the corpus forms a universal dictionary. Whether a word is present in a document is a significant piece of information. However, additional information exists in the document that could be desirable to capture to help us with our document processing problem. Apte et al [13] (like many others) captures the frequency of word occurrence which has produced their good results for document categorisation. Finch [5] has developed an approach to capture information on the sequence of words as they occur in documents. We continue this theme (of extracting more information) by capturing and using contextual dependencies between words. Our thesis in document representation is that words by themselves do not mean very much. Their meaning arises when viewed in some background context that adds information. We believe that capturing and using this background contextual information would be beneficial.

A second benefit of capturing and using contextual dependencies is that it acts as a natural feature selection technique, reducing the dimensionality and size of the feature space. The feature space is a space containing a dimension for each word in the universal dictionary and denominations for each possible value for the dimension. A binary predicate representation will only have two values for each dimension. Representing documents by having a predicate for each word in the universal dictionary is not advisable. The size of the feature space, for a reasonable sized corpus would be too large to perform any computation, such as filtering, efficiently [6]. To reduce the dimensionality of the feature space, feature selection and reparameterisation techniques have been used [6]. Feature selection (removing of unimportant words) and reparamterisation (transformation of feature values) techniques are applied to each document to generate a representation vector. Feature selection techniques have a rich history in statistical pattern recognition [7] and machine learning [8] and their use in document categorisation and filtering is not surprising. Many machine learning techniques for categorisation were born from the fields of mathematical and numerical taxonomy [9]. These fields dealt primarily with the construction of taxonomies for sub-species of flora and fauna [9]. The entities to be classified were instances of a specie whilst the features measure the physical characteristics such as skull width or number of leaves. There exists little background knowledge to determine which of the features were more important than another nor how to transform them. The creation of feature selection techniques helped to overcome this. However, textual documents are quite different from these domains. Words are related to other words in a specific domain and together form, describe and relate to concepts. The words “precision” and “recall” are related in the domain of information retrieval (“IR”). In this context they are used to measure performance in IR. By making use of these dependencies we can replace the words contained in the document, ideally, with a group of words describing the concepts they represent. We have experimentally shown that our approach reduces the dimensionality of the feature space.

We have two arguments in this paper. Firstly making use of contextual dependencies between words in document representation captures important information necessary to work on document processing problems. Secondly, whilst feature selection techniques from numerical taxonomy and machine learning have produced good results and have a place in text categorisation, use of background knowledge can be used to replace and/or supplement them. We have demonstrated this in document categorisation trials on the Reuters-22173 standard document collection. Our results indicate performance comparable to the best published results for the test set [13].

2. Contextual Document Representation

In this section we will describe our approach to capture and use the contextual dependencies between terms. In our studies we have limited the terms to being single words. We first perform statistical analysis of the corpus to extract contextual dependencies between terms. We use these dependencies to build a spreading activation network. The network is then used to process each document to generate a document representation that contains contextual information. In the following subsections we detail the spreading activation technique, our contextual variation of it and an example of the output.

2.1 Spreading activation

Spreading activation is one of the connectionist approaches that is in use in information retrieval (IR) to improve the quality of document processing. The connectionist approaches [10] (neural networks, spreading activation, associative networks etc.) try to model the associative processes that are believed to occur in the human brain. Salton and Buckley [11] provide a good description of spreading activation and an experimental evaluation of one of its variations against other IR methods. We will give a brief description of the simplest variation of the method.

The aim of spreading activation is to determine a set of related terms to an initial set of terms. The terms may be words, word pairs or even complete phrases. Spreading activation uses an activation network consisting of nodes, corresponding to terms, connected by directed weighted links. An initial activation weight is placed on the nodes corresponding to the terms in the document. With each iteration, activation spreads through connections, affecting other nodes. If a_j is the original activation weight of node j, and w_{ij} is the link weight between nodes i and j, representing the influence of node j on node i, the new activation weight a_i’ of node i may then be computed [11] as:

$$W_i = \sum_{\substack{connected \\ nodes}} a_j w_{ij} \qquad (1)$$

$$a_i' = f(W_i)$$

The nodes that accumulate the largest a_i', are selected for the result set. The resultant set of terms can be used to supplement query terms or document terms to improve IR recall. So if a dependency between "precision" and "recall" were to exist and the query or document contained the term "precision" we could add the term "recall". Although the ideas behind these methods are intuitive, "it is fair to say that completely satisfactory models have never been designed" [11]. The spreading activation method improves recall, but causes lost of precision as a result of adding of irrelevant terms to a document or a query.

There are many variations to improve the precision of spreading activation, including use of relevance feedback to adjust the links' weights [12]. In our opinion, one of the main drawbacks of the spreading activation method is that no use of the context that the dependent terms appear in is made. Terms are rarely related unconditionally but usually in some context. The terms, "precision" and "recall" are related in the context of information retrieval, it is doubtful they are related outside the domain.

2.2 Our method

The Idea:

To make use of the context, we introduce a simplistic model of it: a set of terms that are active on a given iteration.

A reasonable weight to place on a link in an activation network is the conditional probability of t_i occurring given t_j occurs, that is, $w_{ij} = p(t_i \mid t_j)$. We approximate this probability by statistical analysis of a large text collection. In our experiments the events t_i and t_j were the occurrence of a specific term in an article or sentence, best results were achieved by using the later. However, we would like to make use of the context in which terms appear in. That is

$$w_{i,j,context} = p(t_i | t_j, context) \qquad (2)$$

$p(t_i | t_j, \text{context}) = p(t_i | t_j, c_1, c_2, \ldots, c_N)$ where c_k is true if term t_k is present in the context, and false otherwise. N is the total number of terms in context.

In our experiments, we arbitrarily limited size of the context to 1 to make it computationally feasible to experiment with reasonably large collections. This gives us:

$$w_{ijk} = p(t_i | t_j, t_k)\, 0 \qquad (3)$$

This can be translated in our "precision" and "recall" example as, the probability that the term "precision" occurs in a sentence given that "recall" and "IR" also occur in the sentence. There are still N^3 possible term combinations, where N is the number of unique terms in the corpus (for non-trivial document collections N is approximately 10^5-10^6). However, we include in the network only those connections where t_j is related to t_i in context of t_k. This implicitly means that strength of this relation is different outside the context of t_k. We have used

$$p(t_i | t_j, t_k) - p(t_i | t_j, \bar{t}_k) \tag{4}$$

as a numerical measure of contextual dependency of terms t_i and t_j in the context of t_k. This is the probability of co-occurrence of terms t_i and t_j in context of t_k minus the probability of their co-occurrence outside the context of t_k. Values above a given threshold indicate a relationship between t_i and t_j that t_k influences.

Nodes in our activation network may be connected by multiple links, each link corresponding to a relationship for a given contextual term. Therefor the weights of links dynamically change depending on the current context:

$$W_i = \sum_{\substack{connected \\ nodes}} a_j v_{ijk} \tag{5}$$

$$v_{ijk} = w_{ijk} \, sign(a_k)$$

$$a_i' = f(W_i)$$

As a result, we have a network where presence or absence (zero weight) of a link depends on the current context.

Underlying Basis:

Our activation network has a variable topology, where the topology depends on the contexts. This allows non-linear effects in the network. We hoped that this would give better performance than other spreading activation methods for at least two reasons:

1. The network and propagation technique maintain that links outside the context will not be activated. "Recall" and "precision" are linked in the context of "IR". But "recall" may be linked to other words though not in the context of "IR". The nodes representing these words will not be activated so long as the context is "IR".
2. From our computations of probabilities, we observed that the strength of dependence between terms is two degrees of magnitude stronger when a single context was in place than when it was not. Notably we found that :

   $\max_{j,k} p(t_i | t_j, t_k) \sim 100 . \max_i p(t_i | t_j)$

Our method, it would appear, is based on stronger dependencies than the traditional spreading activation approach.

Costs:

To derive approximations for probabilities from a large text collection a considerable number of calculations are required.

A number of techniques and considerations allow us to perform the computations for relatively large collections (tens of thousands of documents) in a reasonable time. Discussion of these techniques and considerations is outside of scope of this article. It takes from 3 hours to a couple of days of computing time on a Pentium 120 MHz processor to complete the computations for a document set of approximately 15,000 articles each on average 200 words long.

However, once the network weights have been obtained, the document processing does not take long. It takes less than one second to construct a representation for a 2-5K sized document.

As the network only has to be built once per collection, the computational cost is acceptable. We believe that the influence of a collection specific context dependencies is reasonably weak and it is possible to use a network, built from one collection, to process another collection from a similar domain with an insignificant drop in precision.

2.3 Example

The following is the input and output of our system. The input document is a Reuters' newsfeed article from the test set described later in the paper. The input document shown in figure 1 gives the output from the contextual spreading activation network in figure 2.

swedish industrial production rises sharply stockholm april swedish industrial production rose pct in february after a pct fall in january showing a pct rise over february and reaching its highest level ever the central bureau of statistics said the rise reflected recovery in almost all sectors after an exceptionally cold spell in january the bureau said adding that the highest rises were seen in the forest chemical and metal industries reuter

Figure 1: Reuters's article converted to lower case with numbers removed.

rise_4 figu[re]_3 janu[ary]_3 statistic[s]_3 pct_3 (percentage) industr[ial]_3 produc[tion]_2 febru[ary]_2 swed[ish]_2 bureau_2 december_1 fell_1 indict[ion]_1 season_1 mone[y]_1 consumer_1 compar[e]_1 adjust_1 surplus_1 departm[ent]_1 show_1 revis[e]_1 prev[ious]_1 adding_1, rose_1

Figure 2: Output of processing article through activation network.

The input document is presented to the network as a list of stemmed word frequencies with all numbers removed. The output of the network consists of a word stem and its strength.

The output covers the article's topic (industrial production) and technique (percentage, figure, statistics, january, february) with a high weight. However in addition to these terms are words which do not appear in this particular document but are usually found in the context of the topic (industrial price index) such as fell, season[ally] adjust[ed] figures, previous (comparing to previous results), figure, December (the second previous month before the last result).

3. Applications to Document Classification

We have applied our document representation technique to the problem of document categorisation. We will describe our experiments.

The data for the experiments was drawn from the text categorisation test collection Reuters-22173. This collection was obtained by anonymous ftp from `/pub/reuters` on `ciir-ftp.cs.umass.edu` and consists of articles drawn from the Reuters newswire in 1987. A detailed description of this collection exists [3].

We use the 14704/6746 split of documents into training and test cases respectively as have other studies by Lewis [3], Finch [5] and Apte et al [13]. We remove all documents with no assigned categories as have Finch and Apte et al. The documents are divided in this partition according to date, the documents in the 14704 split (early partition) occur on or before 7 April 1987, and the documents in the 6746 split (late partition) occur from 8 April 1987 onwards. Removing the unclassified documents from the test set resulted in 3301 documents[1].

A spreading activation network using our approach was built from the training set, although it could have been built from the test set or both sets combined. No knowledge of the categories that documents belong to were used. Some of the parameters in our experiments would appear to be arbitrary but are in fact the results of much trial and error.

All the documents containing a category in the training and test sets were processed through this network to generate a representation. Each document was represented by the twenty most highest strength words and strength from the network and the five most frequent words and their frequencies in the document.

[1] Curiously, we found the number of documents left in the test set after this operation was 3301. This differs from the number reported (3672) by Apte et al [13] apparently generated by the same operation. A possible source of this difference may be the large number of documents given incorrect classifications by Reuters. While there are officially only 135 topic categories, more than 201 actual topic categories have been assigned to the Reuters documents—the manual classifiers it would appear occasionally assigned categories which do not officially exist. In our experiments, we have only retained documents assigned to one or more of the 'proper' 135 categories. It is possible that Apte et al included all documents given topic categories, even unofficial categories, and thus came up with a larger number of documents. This was confirmed by email from Apte as a possible source of the discrepancy.

Both singletons (words appearing only in one document), and stopwords (common words) were filtered out. If a word was selected by the network and was also one of the frequent words, the strength given to the word is the frequency count. Figure 2 is an example of a document representation of words and associated strengths.

Inspection of the results of this process show that the strengths for words were rarely distributed evenly across the range of possible values, but tended to cluster in bands. Analysis of the initial output from the categorisation experiments showed that often classes were incorrectly split due to their documents having different values for the same feature within a band. Better performance was achieved by converting all values for an attribute which fell within a band to the same value. This is an example of simple feature reparameterising.

We used a simple Bayesian classifier for our experiments. A priori probabilities were calculated from the training set. The output of the classification process is a probability that the document belongs to a class. We assigned documents to the most probable class(es).

Using this approach we achieved a precision recall break even point of 84% which is comparable to the good results of 80.5% achieved by Apte [13].

4. Discussion and Future Work

While our results are good there is a time and space cost associated with obtaining them though we believe these costs are not excessive. Generating the spreading activation network from the training set (nearly 12 megabytes of text) took approximately eight hours of processing time on a Pentium 120MhZ machine. The generated network size was approximately 2 megabytes whilst the word lookup index was a further half a megabyte. We emphasise that this process only needs to be completed once and the processing time is dependent on the word dependencies in the training set. It takes on average less than a second to generate a description of a two hundred word article from the network and approximately 2 hours to generate the representations for the training and test sets.

Our technique differs from traditional spreading activation in two primary ways. We consider contextual dependencies instead of conditional dependence and we use these dependencies to mostly replace the document terms in the representation instead of supplementing them. In our categorisation experiments only the five most frequent words in the document were retained in the representation, the remaining twenty were obtained from the activation network.

We feel our technique is beneficial for three reasons which are related. Firstly, by making use of the contextual dependencies between words, a document can be represented by a set of words which may not even occur in it. However these words have been shown to occur with the words that are in the documents. Secondly, the dimensionality of the feature space, size of the universal dictionary and the number of unique words used to represent the documents is quite small, this reduces computation time. Finally we partially remove linguistic style qualities in the text (hopefully) leaving only information relating to its topic of content. We will now discuss the last two issues in detail.

Our representation technique generates a small dimensional feature space. In our experiments the number of dimensions in the feature space, which is equal to the number of unique words used to represent the documents is 3279. Each document is represented by between twenty and twenty five words. The dimensionality of the search space using a document representation of the ten most frequently occurring words (half as many as our technique) in each document (removing singletons and stop words) resulted in nearly twice as many (6258) unique words.

We feel that one reason for our good results are due to the removing of the linguistic style component from the document representation. Normally people are interested in classifying articles by the topic of their contents only. Using only the words contained in the document to represent it introduces the author's style into the representation. If an author has a preference for particular terminology or writing style this introduces a bias which detracts from the classification aim of topic categorisation. Using the contexts which the words in the document appear in, rather than the words themselves to represent the document partially overcomes these problems.

It is difficult to compare our results with the studies completed by Finch [5] and Apte et al [13] as there seems to be many ways in handling the Reuters' test set. We have removed unclassified articles as has Apte et al but this has resulted in a different sized training set. This is most likely due to additional undocumented categories which we have removed but Apte et al most likely hasn't. There are 300 such articles accounting for 10% of the test set.

Contextual dependencies between words changes over time and domain. A dependency between say "Thatcher" and "prime" in the context of "minister" is not longer valid as Margaret Thatcher is no longer the prime minister of Britain. Similarly contextual word dependencies in the domain of say science is not applicable to dependencies in literature. The Reuters' test set is about financial news stories written in 1987. If were to feed a news article on sport written in 1996 into the contextual network a document representation would be produced, though it would be not very useful. Knowing when to identify if a contextual network is out of date or not relevant with respect to a document being processed, needs to be studied. Making use of the size and number of node activations could be a good starting point.

5. Conclusion

We have applied the well known principal that humans use background knowledge to solve problems to generate document representations for document processing problems. We capture the contextual dependencies between words for a corpus in a contextual variation of an activation network and apply spreading activation to generate document representations. This approach introduces into the document representation, words which are not in the document themselves. These words, however, have been shown to occur in the context of the words contained in the document. The approach acts as a natural feature selection technique, reducing the dimensionality of the feature space. We feel that our method removes some of the linguistic style bias which can cause problems for topic only categorisation. This is exemplified by our small feature space. We have demonstrated our approach in document categorisation problems using the Reuters's document test set to achieve comparable results to the best known published [13].

References

1. Winston, P.H., Artificial Intelligence 2nd edition, Addison-Wesley (1984) 21-23
2. Houston, J.P., Fundamentals of Learning and Memory 3rd Edition, HJB Inc. (1986) 353-355
3. Lewis, D.D., Representation and Learning In Information Retrieval, Ph.D. Thesis, Department of Computer and Information Science, University of Massachusetts (1992).
4. Salton, G., McGill, M.J., Introduction to Modern Information Retrieval, McGraw-Hill, NY (1983)
5. Finch, S., Partial Orders for Document Representation: A New Methodology for Combining Document Features, ACM SIGIR (1995) 264-272
6. Schutze, H., Hull, D., Pedersen, J., A Comparison of Classifiers and Document Representations for the Routing Problem, ACM SIGIR (1995) 229-237
7. Fukanaka, K., Statistical Pattern Recognition, Prentice Hall (1992)
8. Michalski, R. S., Carbonell, J., Mitchell. T.M., editors. Machine Learning. An Artificial Intelligence Approach. Volume II, Morgan Kaufmann Los Altos CA (1986).
9. Jardine, N., Sibson, R., Mathematical Taxonomy, London, New York, Wiley (1971)
10. Doszkocs, T. E., Reggia, J., Lin, X., Connectionist models and information retrieval. Annual Review of Information Science and Technology (ARIST), (1990) 25:209-260,
11. Salton, G., Buckley, C., On the Use of Spreading Activation Methods in Automatic Information Retrieval, Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1988)
12. Belew, R.K., Adaptive Information Retrieval: Using a connectionist representation to retrieve and learn about documents, Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1989).

13. Apte, C., Damerau, F., Weiss, S., Automated Learning of Decision Rules For Text Categorization, ACM Transactions on Information Systems Vol 12 No. 3 July (1994) 223-251.