PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 48, NO. 3–4, PP. 165–182 (2004)

USING CONCEPT RELATIONSHIPS TO IMPROVE DOCUMENT CATEGORIZATION

Péter SCHÖNHOFEN and Hassan CHARAF

Department of Automation and Applied Informatics
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Fax: (+36) 14632871, Tel.: (+36) 14632870
e-mail: schonhofen, hassan@avalon.aut.bme.hu

Received: September 24, 2003

Abstract

In the information age we much depend on our ability to find information hidden in mostly unstructured and textual documents. This article proposes a simple method in which (as an addition to existing systems) categorization accuracy can be improved, compared to traditional techniques, without requiring any time-consuming or language-dependent computation. That result is achieved by exploiting properties observed in the entire document collection as opposed to individual documents, which may also be regarded as a construction of an approximate concept network (measuring semantic distances). These properties are sufficiently simple to avoid entailing massive computations; however, they try to capture the most fundamental relationships between words or concepts. Experiments performed on the Reuters-21578 news article collections were evaluated using a set of simple measurements estimating clustering efficiency, and in addition by Cluto, a widely used document clustering software. Results show a 5–10% improvement in clustering quality over traditional tf (term frequency) or $tf \times idf$ (term frequency-inverse document frequency) based clustering.

Keywords: document categorization, statistical analysis, document clustering.

1. Introduction

As computing becomes more and more pervasive, the amount of electronically stored information grows at an enormous rate. This, accompanied by the fast changing technological environment and the increasingly wide range of users, has led to the current situation, where valuable information is buried in collections having ad-hoc or unstructured formats. Most of this information resides in documents written in natural language: World-Wide-Web pages, corporate e-mail archives, digital library documents and so on.

Although natural language is suitable for human understanding, when the information reaches a critical mass, the cost of data categorization and selection performed by human beings becomes unbearably high. However, algorithms performing sophisticated natural language analysis are either too computationally intense or too restricted in their aptitude to be practically usable. Instead, we have to settle for inaccurate but fast methods.

In this article we focus on document retrieval, that is, finding the words (or concepts), which most effectively represent a given document during the categorization of a specific document collection (SEBASTIANI [12] and YANG [15] provide an overview of automated text categorization). Such a method should posses the following qualities:

- Selection of representative words (or concepts) should be fast and its time should grow only linearly with the collection size.
- The amount of information stored about each document in order to facilitate word (or concept) selection should be minimal.
- The selected words (or concepts) should represent the original document during classification in an accurate and effective way.

The fundamental idea is that documents should not be regarded as independent entities during selection of representative concepts, but rather as integral parts of the same collection. We mediate this 'union' toward the documents through a primitive concept network describing general-specific and co-occurrence relationships between words.

2. Employed Corpus

The method presented in the following sections was tried out and evaluated against traditional techniques using the Reuters-21578 corpus (see Reuters in the references). This collection contains short news items about various topics (mainly in the domain of politics and economics) written in English, and many of these documents were assigned by humans to one or more of 135 categories – making it an ideal base on which to test retrieval methods. Low-quality articles and category assignments were clearly marked as such in the corpus; however, not all documents and categories present in the collection were suitable for processing: categories containing fewer than 10 or more than 200 articles were discarded, and only articles assigned to at least one category and consisting of 50–300 words were kept. These steps were necessary to limit category and document sizes to a reasonable range, thus providing a document collection of roughly homogeneous properties and categories comparable to each other. (The Reuters corpus contains extremely large and small categories, some consisting of as many as 4000 documents, others being assigned to a single document.) As a consequence, experiments were performed on 1833 documents.

As a pre-processing step, the documents were parsed to isolate words, numbers, sentences and other lexical formations (abbreviations, signs, type codes and so on), using WordNet (see WordNet in the references) for both stemming and stop word removal, this latter slightly modified manually by the authors to adjust to the unique nature of the applied corpus. Although there are more aggressive methods to automatically detect and remove redundant words specific to the processed corpus (see for instance YANG [14]), we opted not to use them, instead relying entirely on our algorithm, whose aim is to recognize irrelevant words, a broader objective. Similarly, WordNet could have played a larger role, providing semantical information about words and word pairs, as was the case in Rodríguez [10], for example. However, in order to make our method as language independent as possible, we avoided the use of more advanced natural language processing techniques (such as measuring semantical distance between words), which may not be available for every language.

The lexical elements had the following distribution:

stop words:	48%
non-words:	8%
valid words:	44%

Although fewer than half of all lexical elements encountered were kept for further analysis, the number of words remained still relatively high at 206,526 (112.67 words for each document in average), providing a sufficient amount of data for basic statistical analysis. Out of the 1,833 documents, 440 had more than one category assigned to them, resulting in slightly overlapping categorization.

The fact that experiments were carried out on documents written in English does not mean that results does not apply to other languages as well. It will be clear from the detailed description of our proposed method that we carefully avoided language specific processing. Another concern, which should be addressed, is the particular nature of the documents – one might wonder whether our method would behave the same way when applied not to relatively short news articles but instead to lengthy technical documents, for example. But longer texts containing fewer unique proper names and more occurrences of the same terms would strengthen the performance, as the document model would be based on a much broader foundation.

3. Document Model

To simplify and speed-up further analysis, the document collection is replaced by four measurement sets, called the document model, which are summarized in the table below (frequency-based feature extraction approaches were followed also by Aizawa [1] and Debole [3], among others, although in different frameworks):

global frequency	global context frequency
local frequency	local context frequency

Before the exact meaning of these measurements is presented, the concept of ‘context’ should be defined. Context of a given word having a specified location in a document is the set of words occurring near that word, in the same sentence. More precisely, context consists of the preceding and subsequent R words (if they exist), where R denotes the so-called context range. However, because handling word-sets is tedious and uncomfortable in practical data processing applications, instead, we will employ word-pairs, where the second word is present in the context of the first one – thus the context of a given word with $R = 3$ translates to at most six word-pairs.

Note that word pairs are not bigrams (described, for instance, in EVANS and ZHAI [4]), as the constituting words are not necessarily adjacent and the pair itself is not always of high frequency. (If the scope of documents in the corpus was more restricted, thus carrying a more limited vocabulary and set of phrases, the use of such 'loose' N-grams (where $N > 2$) would have seemed reasonable.) By allowing intervening words between the two members of a word pair, we make possible the inclusion of more general concepts in the document model, such as 'performance analysis', which might not occur as a direct technical term in the document, but rather as scattered words in the sentence: 'we should analyse the overall system performance'. A disadvantage of this approach is the introduction of non-related words as pairs ('overall system' in our example), but their frequency will be sufficiently low to exclude them from further processing. Finally, of course word-pairs are formed only after document pre-processing has taken place (described in the prior section), and thus stop words do not participate in contexts, not limiting its scope.

Names of the four measurement sets allude to their function: global data describe the entire collection and local data correspond to a given document; likewise, context frequencies pertain to word-pairs while regular frequencies characterize individual words. They are different aspects of the same phenomenon.

Now let us introduce the document model. The most extensive and detailed one of the four measurement sets is the local context frequency, specifying the number of times a word-pair occurs in a document (a similar approach is presented in MATSUO and ISHIZUKA [8]). Two remarks should be made: First, if a word occurs in the context of itself, the resulting word-pair is ignored since it does not carry valuable information. Secondly , there is symmetry between word-pairs, since the existence of a word-pair $A - B$ implies the presence of $B - A$; still, due to the fact that a word occurring multiple times in context of an other word generates only one word-pair 'instance', their associated frequencies would not be necessarily the same. Consider the following theoretical word sequence:

$$abcbd$$

Here the pair $b - c$ occurs twice, while its counterpart, $c - b$, does only once, because the two bs in the context of c are not distinguished from each other, thus resulting in a single word-pair – as opposed to $b - c$, which means two contexts of b, containing the same instance of c. A possible and also reasonable solution is to assign the minimum of these two frequencies to both word-pairs, hence making them equivalent.

The higher the *local context frequency*, the closer the relation is (possibly conceptual) between the two words forming the pair. When the context range is zero, word-pairs are limited to adjacent words, while larger context ranges facilitate the recognition of more implicit concepts, but only in exchange for blurring word locality and for needing more storage capacity, though latter increases only linearly with R in the worst case, and hence does not burden significantly further computations. We set R to 3 during the experiments measuring the performance of our proposed methods (described later), a sufficiently large value to cover all valuable concepts.

As its name tells, *global context frequency* is simply the sum of its local counterpart across the entire document collection; that is, it specifies how many times a given word-pair occurs in any document. Word co-occurrence is more reliably indicated by a high global context frequency, but only with the reservation that large individual documents may have a distorting effect.

As opposed to context frequencies, *local* and *global frequencies* merely record the number a given word occurs in a document (the same as tf, term frequency) or in any document located in the collection, respectively. Note that the local frequency of a word A cannot be exactly calculated by summing all local context frequencies involving A in the given document, as we ignored multiple occurrences in the same context.

In order to prevent disproportionate influence of extremely frequent words or large documents on frequency data, they should be normalized, thus the formulae used to calculate the various measurements become as follows:

Global frequency:

$$F'_w = \frac{F_w}{\sum_{w^*} F_{w^*}}$$

Global context frequency:

$$C'_{w_1 w_2} = \frac{C_{w_1 w_2}}{\sqrt{F_{w_1} F_{w_2}}}$$

Local frequency:

$$L'_{wd} = \frac{L_{wd}}{\sum_{d^*} L_{wd^*}}$$

Local context frequency:

$$D'_{w_1 w_2 d} = \frac{D_{w_1 w_2 d}}{\sqrt{L_{w_1 d} L_{w_2 d}}},$$

where X' means the normalized value of measurement X, while w, w^*, d and d^* designate words and documents, respectively (in case of global and local frequencies, the sum in the denominator is performed over all words or documents in the collection). It may seem unconventional that local frequency is normalized by the sum of local frequencies of the given word across the whole document collection, and not by the sum of local frequencies of words present in the given document. The explanation is that we regard local frequency more as the property of words than as of documents.

The four document model components listed above describe both the document collection and its members in sufficient detail for our subsequent analysis, yet considerably reduce the required storage (inclusion of another data would also be possible, see for instance GAWRYSIAK et al. [5]; for a more semantic-focused approach refer to CAI and HOFMANN [2]). Assuming a context range of 3, for the 1,833 documents, this means the following:

global frequency:	10,877 words
global context frequency:	332,976 word-pairs
local frequency:	137,635 words
local context frequency:	549,926 word-pairs

Compared to the traditional indexing technique, where for each word references to the documents it occurs in are recorded, this document model yields far more data. In our case, indexing would require 148,512 storage cells (sum of global and local frequencies) in contrast to the 1,031,414 cells mandated by the document model, a 694% increase! (Frequency values, word and document references were regarded as a single cell.)

Obviously, this is untenable. Hence we remove all data elements which do not have analysis value – in practice, this means that words or word-pairs occurring only once in the processed subset of the collection (that is, in the 1,833 selected documents) are omitted from both local and global measurement sets. The required storage capacity for each measurement set is now:

global frequency:	7,047 words	(64%)
global context frequency:	88,298 word-pairs	(26%)
local frequency:	133,805 words	(97%)
local context frequency:	295,102 word-pairs	(53%)

At each item we show that the current storage need is what percentage of the previous one. The total amount of storage cells is now 524,252 (50%), still larger than that required for traditional indexing (198%), but not by so much.

4. Selecting Concepts

After the document model has been built, the words – or generally speaking, the concepts – typical of each document must be selected, which will represent these documents during the retrieval process. Actually, representative words play two distinct roles: some of them determine the category the document pertains to, while others distinguish it from other documents in the same category. (Obviously, a specific word can be a ‘differentiator’ when there are few categories, but might be a category ‘designator’ when the number of categories is larger.)

Here our primary concern is category ‘designators’, so the question is: Which concepts are characteristic of a document? Which words are central to its content? A reasonable answer is that concepts being the most interwoven with the fabric of the document, that is, those most tightly coupled with the other words present, regarding various concept relationships.

The key in the above statement is ‘concept relationships’, which derives either from meaning (general-specific, part-whole etc.) or from language usage (multi-word technical term, idiom etc.). Relationships also have strength: the general-specific connection between the words ‘furniture’ and ‘table’ is clearly stronger than between ‘object’ and ‘table’, as latter relation goes through more intermediate concepts (among them ‘furniture’).

When we examine how strongly a specific word relates to other words mentioned in the same document, there are several factors which may or may not be taken into account:

- number of words with which the given word has any relationship
- average or accumulated strength of these relationships
- evenness of these relationship strengths (measured by standard deviation)
- completeness of relationships[1]
- type of relationship (for example, a general-specific relation may have greater impact than a phrase connection)

Selecting representative words is not enough, we should evaluate how this selection performs against categorization (and not direct retrieval, since we said we will focus on category indicators). The evaluation was performed in two parts: (1) we have calculated the value of six measurements characterizing how well the representative words would help document clustering; and (2) we actually clustered the documents based solely on the representative words using Cluto 2.1[2], a free available document clustering software (see KARYPIS [7] for a detailed description).

The evaluation measurements are category-based, and hence have to be averaged across categories after they are computed – a plus sign indicates that the corresponding measurement value should be as high as possible for a high-quality clustering, the minus sign that it should be as low as possible.

- Vocabulary-based measurements (the most rough):
  - number of different words present in documents pertaining to C (*width*)
  - number of different words present in documents pertaining to C which also occur in documents assigned to other categories (*overlap*)
- Distribution-based measurements:
  - for the different words present in documents pertaining to C, average percentage of these same documents containing them (*coherence*)
  - for the different words present in documents pertaining to C, sum of the number of documents assigned to other categories containing each word, normalized by the number of different words occurring in C (*blur*)
- Similarity-based measurements (the most sophisticated):

[1] Assume that there is a relationship chain $w_1 - w_2 - \ldots - w_n$ of the same type (for instance from ‘object’ to ‘furniture’ to ‘table’), and that words w_1 and w_n occur in the document in question – completeness is then defined as the percentage of words $w_2, w_3, \ldots, w_{n-1}$ also present in the document. Obviously, this definition is valid only for general-specific and part-whole relationships; in case of a multi-word term $w_1 - w_2$, completeness may mean the percentage of words from the domain described by w_1 and w_2 also occurring in pair with either w_1 or w_2.

[2] The clustering solution was computed using repeated bisection followed by global optimization, and for evaluation we chose the $G1'$ criterion function.

- maximal distance[3] between any two different documents in C (*diameter*)
- maximal similarity[4] between a document pertaining to C and an other assigned to another category (*separation*); no document is compared to itself, even when it is assigned to more than one category.

Where C designates the category we want to measure; if a document pertains to more than one category, the document will be involved when computing the quality measurements of each category. We could have adopted measurements employed in the information retrieval community, such as precision-recall or entropy-purity. However, the measurements presented above seemed more appropriate for our task, because they are easy to compute, and in the same time they characterize several aspects of the attainable categorization precision.

5. Selection

One thinks that the best way to generate relationships and compute their strength would be to employ a rich thesaurus database, such as WordNet (see SCOTT and MATWIN [11] for an application of WordNet to improve document classification). However, for the 1,833 documents processed, the percentage of word-pairs for which a WordNet-defined relation exists is about 0.45% — far too low to yield a sufficient number of representative word candidates. The likely cause is partly the high number of proper names in the news articles comprising the collection, partly their terse and stylistically rich wording.

Instead, we have to approximate concept relationships using the document model computed from word and word pair occurrence counts; in order to make the four measurements constituting the document model comparable to each other, we scaled them to the [0, 1] interval: global frequency data globally, while local frequency data in each document. Note that this second normalization was performed independently of the first one (as described in Section 3), their objective was entirely different.

We tried out five different approaches to estimate which words are the most central to the topic of individual documents, and therefore, the most suitable to represent documents during clustering. In each case we heuristically constructed a formula to grasp the essence of the given approach, then used this formula to compute the rank of words present in a document. Higher values mean better (and smaller) rank for the word; when the formula gives the same value for two different words, they naturally receive the same rank, but in exchange no word will be assigned to the next rank.

[3] Distance between two documents is calculated as the number of different words present in only one of them.

[4] Similarity is the opposite of distance: we define it as the number of different words present in both documents.

In the first approach, called *simple selection,* we selected or rejected a word as representative of the document based on how many relationships it formed with the other words occurring in the same document – giving the following scoring formula:

$$r_{wd} = |S_{wd}|\,,$$

where r_{wd} stands for the score of word w in document d (on which word ranking will be based), S_{wd} denotes the set of words in relationship with word w in document d, and $||$ means the set size.

In *weighted selection*, the second approach, we did not focus on the number of relationships, but rather on their strength: it was assumed that words conceptually connected to a large number of other words in the document with weak relations would be more effective representatives than words with stronger relationships (presumably the attribute of common usage words). The employed formula was:

$$r_{wd} = \sum_{w^* \in S_{wd}} \frac{1}{D'_{ww^*d}},$$

where w^* denotes a word in relation with the examined word w, and D stands for local context frequency, as defined in the previous section.

In the third case, *evenness selection*, words whose relationships with the other words in a given document have approximately the same strength (and at the same time are weak) are preferred to words with relations of widely varying intensity. Our reasoning was that if a word is discussed in a detailed manner, presenting all its aspects in a wide range of contexts, it should be central to the document topic. The applied formula was as written below:

$$r_{wd} = e^{\mathrm{Dev}_{[w^* \in S_w]}\left(D'_{ww^*d}\right)} \min_{w^* \in S_w} \left\{D'_{ww^*d}\right\},$$

where Dev denotes standard deviation; all other notations are the same as was described in the previous approaches.

The fourth method, named *combined selection,* merged the three formulae introduced so far – therefore the computation of word rank is slightly different than previously: we now use directly the ranks associated with the selection formulae, and not the formulae themselves. Assuming that the best rank is 0, the formula is:

$$r_{wd} = -\max\left\{s_1;\ \frac{s_2}{3};\ \frac{s_3}{6}\right\},$$

where s_1, s_2 and s_3 designate the ranks word w received from simple, weighted and evenness selections, respectively; their maximal value is taken, so that only words equally excelling in all three aspects get attention. The minus sign is necessary because now a value closer to zero means a word more suitable as a document representative; s_2 and s_3 are divided by 3 and 6, respectively, to reflect their lesser role as compared to s_1.

Finally, *balanced selection* takes all measurements available into consideration about a relationship, preferring locally frequent but globally rare words (characteristic of the examined document) having a weak relation with locally and globally rare words (too specific to represent the document topic). The employed formula is as follows:

$$r_{wd} = L'_w \frac{1}{1 + \ln F'^{-1}_w} \sum_{w^* \in S_{wd}} \left(\frac{1}{D'_{ww^*d}} \frac{1}{L'_{w^*}} \frac{1}{1 + \ln F'^{-1}_{w^*}} \right).$$

We take the logarithm of global measurements, as the distribution of their values strongly tends to zero (addition of 1 is necessary to avoid division by zero, when w or w^* is the most frequent word globally).

In order to compare the performance of these selections to traditional tf (term frequency) and $tf \times idf$ (term frequency-inverse document frequency) methods, the following two additional ranking formulae had to be introduced:

$$\begin{aligned} r^{TF}_{wd} &= L'_{wd} \\ r^{IDF}_{wd} &= L'_{wd} \log \frac{N}{P_w}, \end{aligned}$$

where N is the number of documents in the collection, and P_w denotes the number of documents containing word w.

Note that in each presented approach we estimated concept relationships and their strength by various frequency data, and did not take more sophisticated properties mentioned in Section 4 into account, such as relation type and completeness. However, doing so would have imposed a heavy computational burden on our proposed method, and thus would have made it impracticable in real-word document retrieval situations.

6. Results

Execution and evaluation of the different word selection methods – described in Section 5 – were carried out with parameters varying along two dimensions.

The first parameter determined whether overlapping categories were allowed or not; that is, either all 1,833 documents (from 49 different categories) or only 1,354 documents (from 31 categories) were involved in the experiments. Because the document clustering software could not handle documents pertaining to more than one category, in the first case the category assignment of these documents was reduced to the largest category (those containing the most documents), leaving 39 categories.

The second parameter specified how many words were kept as representatives from each document, or more precisely, what was their maximal allowed rank; our experiments were carried out with maximal ranks of 0, 1, 2, 3, 4 and 9. It was possible that multiple words received the same rank, so the actual number of representatives – called the *selection depth* – often exceeded the specified word count.

In order to see how far the traditional and proposed methods fall from an ideal selection method, the concept of *optimal selection* was introduced: here documents were represented by one or more special words, each corresponding to a category assigned to the given document. Due to its particular nature, selection depth could not be controlled, so when we allowed overlapping categories, optimal selection chose 1.424 words in average for a document, otherwise each was represented by exactly one word.

Now let us see the actual results for the various selections along the previously mentioned dimensions, both through the six measurements characterizing selection quality (described in Section 4) and the *entropy-purity* value computed by Cluto after clustering the documents using solely their representative words. Entropy signifies how evenly the various document categories are distributed in each discovered cluster, while purity specifies whether clusters contain documents mainly from a single category or not. High purity and low entropy values designate a high quality clustering (see Zhao and Karypis [16] for a more detailed description).

Tables 1–2 show the performance of selection based on tf and $tf \times idf$, while *Tables 3–8* contain results achieved by our proposed methods. Data in non-shaded rows refer to the case when overlapping categories were allowed, while shaded rows contain values measured when only single-category documents were processed. The first column of the tables show the maximal rank value allowed for the given selection.

Table 1. Results of tf-based selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	43.04	34.22	5.13	9.61	4.53	1.92	0.334	0.609
1	78.31	64.00	5.50	13.90	8.10	2.90	0.376	0.560
2	114.67	96.71	5.67	17.81	12.18	4.18	0.410	0.519
3	151.51	129.90	5.85	21.25	16.14	6.02	0.444	0.493
4	187.67	162.33	6.00	23.60	19.98	7.98	0.482	0.458
9	396.35	358.02	6.18	35.75	52.33	16.71	0.535	0.386
0	34.10	18.58	6.35	5.30	4.48	1.45	0.272	0.702
1	63.87	37.61	6.45	8.12	8.32	2.23	0.317	0.657
2	94.06	61.23	6.55	11.14	10.87	2.97	0.362	0.610
3	125.77	86.16	6.70	13.53	14.84	3.71	0.412	0.566
4	155.97	109.84	6.75	15.96	18.35	5.10	0.462	0.490
9	358.87	289.03	6.79	25.75	54.13	11.97	0.520	0.436

Table 8 lists selection depths measured at the various methods, in the same format as the previous tables; as it can easily be seen, the method where depth

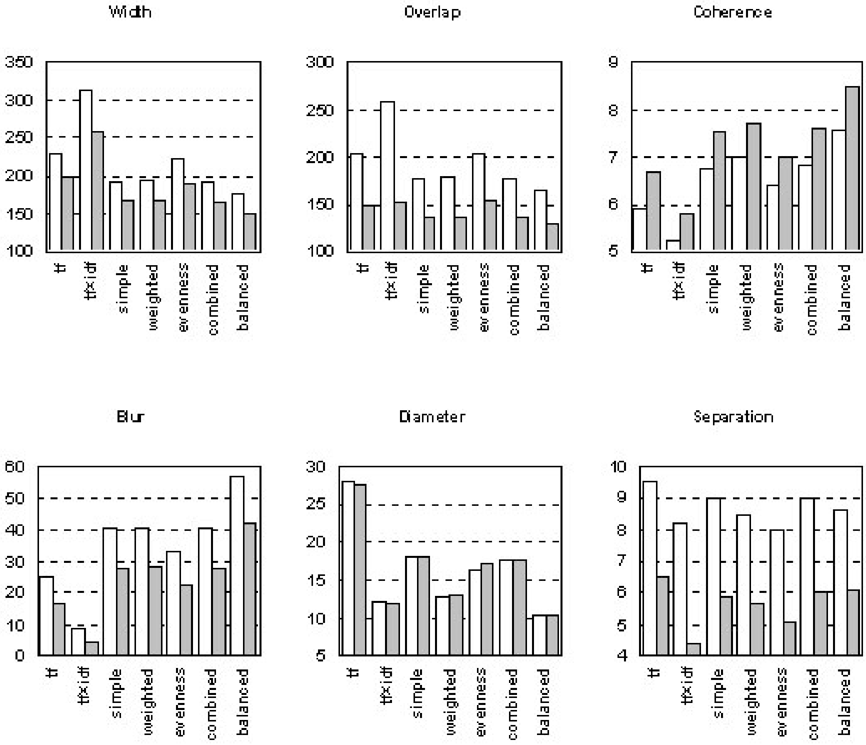

Fig. 1. Measurement values at selection depth 10. White and grey bars show data measured when overlapping categories were allowed or not, respectively.

follows the allowed maximal rank the most closely is the balanced selection – in other words, there is the lowest of the probability of two words receiving the same rank.

Because depth heavily influences clustering quality (more words represent the document better), the different methods cannot be compared properly, unless we can bring measurement values in a way to a common selection depth point. Fortunately, our six evaluation measurements give values growing linearly with depth, while entropy and purity values given by Cluto follow a logarithmic curve, therefore, interpolation is easily carried out. *Fig.* 1 shows the evaluation measurements and *Fig.* 2 compares clustering quality values, as approximated at depth 10. *Fig.*2 includes results for both optimal selection (note that its depth is 1.424 for overlapping categories and for single-category documents) and full text categorization (see bars with label 'full'), when clustering was performed using all words in the documents.

Both *Fig.* 1 and *Fig.* 2 clearly illustrate that clustering can be performed

Table 2. Results of $tf \times idf$-based selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	39.80	25.51	4.57	1.94	1.76	1.04	0.677	0.286
1	72.45	49.67	4.76	2.76	2.84	2.08	0.529	0.378
2	104.76	74.69	4.85	3.57	4.04	2.73	0.495	0.414
3	135.41	100.76	4.96	4.61	5.71	3.37	0.475	0.448
4	166.59	125.98	4.99	5.38	6.98	4.18	0.435	0.482
9	317.59	264.86	5.16	8.35	12.06	8.41	0.390	0.542
0	31.16	7.37	5.33	0.64	1.61	1.07	0.683	0.321
1	57.19	17.97	5.77	1.21	2.71	1.23	0.532	0.395
2	83.10	30.26	5.74	1.60	4.16	1.55	0.493	0.430
3	107.81	46.13	5.79	2.15	5.06	2.06	0.441	0.509
4	133.32	60.58	5.75	2.62	6.52	2.52	0.408	0.538
9	261.55	160.90	5.72	4.62	12.13	4.45	0.336	0.629

Table 3. Results of simple selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	29.63	25.96	6.04	17.32	4.31	1.49	0.526	0.380
1	57.47	51.37	6.27	24.59	6.27	2.73	0.471	0.460
2	84.08	76.27	6.32	28.42	8.78	3.61	0.427	0.520
3	111.27	101.31	6.49	31.68	10.94	4.96	0.392	0.542
4	139.80	127.80	6.52	34.65	13.61	6.16	0.366	0.574
9	252.82	233.67	7.00	46.35	23.41	12.10	0.315	0.631
0	23.45	16.32	7.09	10.38	4.00	1.23	0.521	0.441
1	48.10	36.52	7.18	15.82	6.55	2.32	0.455	0.519
2	70.29	55.52	7.23	18.66	9.06	3.00	0.395	0.580
3	93.58	74.68	7.31	21.72	11.32	3.45	0.353	0.622
4	118.81	95.58	7.32	24.01	13.81	4.03	0.337	0.636
9	222.00	185.42	7.71	32.75	23.32	7.84	0.282	0.682

with more accuracy when processing single-category documents than when we allow overlapping categories; the only measurement which does not reflect this is diameter, possibly because vocabulary of documents belonging to the same category differ as much as those assigned to two different categories (though frequency of differing words may be higher in the latter case).

When considering entropy and purity as sole indicators of clustering efficiency, simple and combined selections emerge as the best approaches, performing 5–10% better than traditional tf and $tf \times idf$ methods, even coming close to levels

Table 4. Results of weighed selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	23.86	20.49	5.99	15.44	1.16	1.00	0.577	0.392
1	44.73	39.55	6.22	22.17	2.55	2.00	0.490	0.433
2	65.22	58.65	6.40	25.38	3.80	2.84	0.444	0.497
3	85.82	77.90	6.49	28.18	4.92	3.61	0.430	0.511
4	103.86	94.27	6.70	30.77	6.10	4.45	0.390	0.548
9	197.24	181.71	6.92	39.75	13.06	8.63	0.351	0.589
0	18.45	12.10	7.05	8.77	1.16	1.00	0.564	0.457
1	35.16	24.97	7.40	13.60	2.48	1.90	0.486	0.462
2	52.68	39.84	7.48	16.65	3.81	2.42	0.407	0.570
3	69.61	54.13	7.44	18.79	5.06	2.94	0.377	0.598
4	86.45	68.10	7.53	20.34	6.23	3.29	0.352	0.623
9	171.13	140.74	7.62	27.54	13.61	5.71	0.290	0.687

Table 5. Results of evenness selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	31.63	25.98	5.04	9.20	2.16	1.00	0.632	0.336
1	55.35	47.43	5.31	13.64	3.08	2.00	0.524	0.403
2	80.63	70.22	5.58	17.12	5.69	2.76	0.485	0.461
3	103.47	92.31	5.74	20.63	7.10	3.53	0.453	0.483
4	126.18	113.37	5.90	23.13	8.67	4.24	0.430	0.504
9	230.86	211.90	6.35	32.65	17.16	8.37	0.356	0.579
0	25.29	14.74	5.97	5.08	2.29	1.00	0.641	0.363
1	44.32	29.52	6.29	8.42	3.23	1.90	0.511	0.454
2	66.03	46.55	6.44	10.61	6.10	2.32	0.466	0.507
3	85.10	63.71	6.60	13.23	7.03	2.84	0.423	0.543
4	104.61	79.97	6.69	15.00	8.87	3.16	0.383	0.592
9	198.81	162.26	6.95	22.29	18.16	5.19	0.328	0.629

produced by full text categorization, though the number of representative words is seven times less.

However, the picture painted by the six evaluation measurements is not so unambiguous since simple and combined selections do not show clear superiority in every aspect. This might be attributed partly to the sophisticated interplay between the various phenomena observed by these measurements – for instance, if we choose very specific and rare words from documents, separation and blur improves while coherence, along with width, will deteriorate; similarly, choosing common, globally

Table 6. Results of combined selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	24.59	21.18	6.02	16.08	2.06	1.00	0.571	0.387
1	49.57	44.59	6.32	23.88	4.55	2.55	0.480	0.446
2	74.57	67.47	6.44	27.94	6.67	3.24	0.431	0.517
3	100.82	91.71	6.57	30.73	8.78	4.57	0.398	0.544
4	129.45	117.98	6.57	33.90	11.82	5.82	0.372	0.573
9	247.53	228.82	6.98	45.41	22.84	11.82	0.336	0.609
0	19.58	13.10	6.87	9.38	2.10	1.00	0.555	0.444
1	40.10	30.16	7.44	15.31	4.39	1.97	0.449	0.519
2	61.68	47.87	7.30	18.13	6.74	2.61	0.395	0.586
3	83.39	65.74	7.41	21.13	8.74	3.19	0.366	0.608
4	110.00	88.26	7.36	23.06	12.19	3.97	0.331	0.637
9	216.77	181.13	7.70	32.04	22.94	7.84	0.286	0.687

Table 7. Results of balanced selection

	Width	Overlap	Coher.	Blur	Diamet.	Separat.	Entr.	Purity
0	23.04	20.65	6.18	24.40	1.04	1.00	0.603	0.364
1	42.14	38.29	6.58	31.11	2.12	2.00	0.517	0.411
2	60.41	55.61	6.66	36.52	3.24	2.90	0.466	0.466
3	78.49	72.76	6.78	39.02	4.14	3.65	0.427	0.504
4	96.04	89.63	6.93	41.84	5.37	4.55	0.395	0.543
9	173.84	162.80	7.50	55.08	10.37	8.57	0.347	0.587
0	18.16	13.32	7.45	15.81	1.06	1.00	0.598	0.421
1	33.77	26.23	7.69	22.59	2.06	1.97	0.514	0.432
2	48.58	39.19	8.01	26.30	3.13	2.55	0.445	0.526
3	63.55	52.42	8.09	28.35	4.10	2.94	0.384	0.584
4	78.77	66.23	8.12	30.00	5.19	3.39	0.356	0.609
9	149.29	127.74	8.37	40.53	10.23	6.06	0.291	0.688

high frequency word leads to good coherence but inferior blur.

Consequently, the particular document retrieval scenario will determine in the end which method produces the best results. If keywords characteristic to a given document have to be selected (where separation and blur are the most important indicators), we would employ tf-based selection, while for topic identification (where coherence and diameter seem to capture the requirements) balanced selection is the most suitable.

Table 8. Selection depth

	tf	$tf \times idf$	Simple	Weight.	Evenn.	Combin.	Balanc.
0	1.47	1.03	1.25	1.00	1.04	1.04	1.00
1	2.90	2.03	2.61	2.01	2.03	2.28	2.00
2	4.42	3.05	3.89	3.03	3.11	3.50	3.00
3	6.05	4.05	5.26	4.04	4.15	4.79	4.00
4	7.69	5.09	6.76	5.06	5.20	6.29	5.01
9	18.20	10.17	13.61	10.24	10.52	13.24	10.01
0	1.47	1.03	1.24	1.00	1.05	1.05	1.00
1	2.91	2.03	2.61	2.01	2.04	2.28	2.00
2	4.44	3.05	3.93	3.03	3.12	3.52	3.00
3	6.05	4.05	5.28	4.05	4.15	4.78	4.00
4	7.68	5.09	6.80	5.06	5.21	6.30	5.01
9	18.70	10.17	13.68	10.27	10.58	13.31	10.01

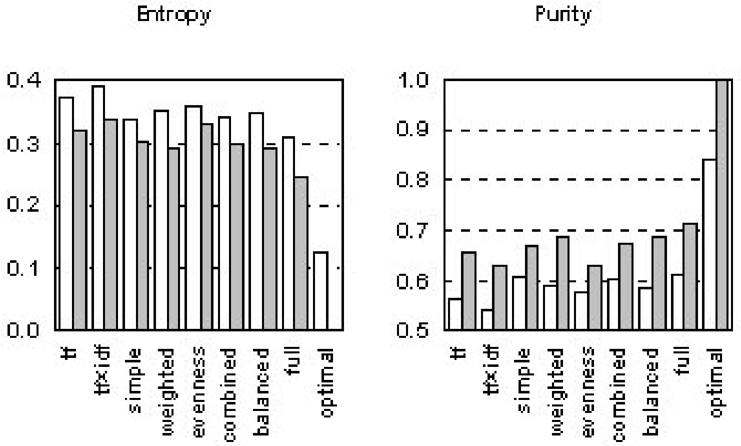

Fig. 2. Clustering quality values at selection depth 10. White and grey bars show data measured when overlapping categories were allowed or not, respectively.

7. Conclusion

The goal of document representatives is twofold: first, their rank values may be a reliable indicator of word relevance in a specific document; and second, by replacing documents with the set of representative words during categorization, the computational effort (in addition to required storage) can be significantly reduced (for an alternative approach, see for example KARYPIS and HAN [7]). Of course, categorizing documents will always be more accurate when processing all words present in the document, but in some situations (for instance when analysing Word-Wide-Web pages as opposed to short abstracts) the reduced quality is acceptable in return for an increased throughput.

In the present article, the authors introduced several methods for selecting words used as document representatives during categorization and clustering, and also various metrics to evaluate them. Results showed that document reduction made categories easier to recognize, but on the other hand, in practical document clustering situations, clustering quality can be improved by 5-10% (measured by entropy and purity) as compared to traditional methods which select words based on term frequency and inverse document frequency. However, the word selection approach suitable for clustering might not be optimal for other document retrieval situations, for instance, when looking for keywords typical of a given document.

Although our experiments were performed on a corpus comprising rather short, English-language documents, and because the proposed methods do not exploit features dependent on either language (such as deep syntactic parsing) or document format (for example recognition of document structure, including sections and paragraphs), their results can be applied to other languages and document collections as well, only the context range, denoted by R, may need adjustment.

References

[1] AIZAWA, A., Linguistic Techniques to Improve the Performance of Automatic Text Categorization, *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, (2001), pp. 307–314.

[2] CAI, L. – HOFMANN, T., Text Categorization by Boosting Automatically Extracted Concepts, *Proceedings of the 26th Annual International ACM SIGIR Conference*, Toronto, Canada, 2003.

[3] DEBOLE, F. – SEBASTIANI, F., Supervised Term Weighting for Automated Text Categorization, *Proceedings of the 18th ACM Symposium on Applied Computing*, Melbourne, Florida, 2003, pp. 784–788.

[4] EVANS, D. A. – ZHAI, C., Noun-Phrase Analysis in Unrestricted Text for Information Retrieval, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 1996, pp. 17–24.

[5] GAWRYSIAK, P. – GANCARZ, L. – OKONIEWSKI, M., Recording Word Position Information for Improved Document Categorization, *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Taipei, Taiwan, 2002.

[6] KARYPIS, G., Cluto: A Clustering Toolkit, *Technical report #02-017*. University of Minnesota, Department of Computer Science, 2002.

[7] KARYPIS, G. – HAN, E.-H., Concept Indexing – A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization, *Technical report #00-016*. University of Minnesota, Department of Computer Science., 2000.

[8] MATSUO, Y. – ISHIZUKA, M., Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, *Proceedings of the 16th International FLAIRS Conference*, St. Augustine, Floridam 2003.

[9] Reuters: The Reuters 21578 Distribution 1.0 collection is available at: http://www.daviddlewis.com/resources/testcollections/reuters21578/

[10] RODRÍGUEZ, M. B. – GÓMEZ-HIDALGO, J. M. – DÍAZ-AGUDO, B., Using Wordnet to Complement Training Information in Text Categorization, *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 1997.

[11] SCOTT, S. – MATWIN, S., Text Classification Using WordNet Hypernyms, *Proceedings of Conference on the Use of WordNet in Natural Language Processing Systems*. Association for Computational Linguistics. Somerset, New Jersey, 1999, pp. 38–44.

[12] SEBASTIANI, F., A Tutorial on Automated Text Categorization. *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, Buenos Aires, Argentina, 1999, pp. 7–35.

[13] WordNet: The WordNet database and its associated programs are available at: http://www.cogsci.princeton.edu/ wn/

[14] YANG, Y. – WILBUR, J., Using Corpus Statistics to Remove Redundant Words in Text Categorization, *Journal of American Society of Information Science*, **47** No. 5 (1996), pp. 357–369.

[15] YANG, Y., An Evaluation of Statistical Approaches to Text Categorization, *Information Retrieval*, **1** No. 1–2 (1999), pp. 69–90.

[16] ZHAO, Y. – KARYPIS, G., Criterion Functions for Document Clustering – Experiments and Analysis, *Technical report #01-40*, University of Minnesota, Department of Computer Science, 2002.