AUTOMATIC RECOGNITION OF KEY-WORDS USING N-GRAMS

Radwan Jalam, Jean-Hugues Chauchat and Jean Dumais

Key words: n-gram, Text representation, Text categorization, Text mining.
COMPSTAT 2004 section: Statistical Data Mining.

Abstract: Documentary research (e.g. bibliography for a thesis or a research project, technological monitoring, etc.) is often based on a few selected key-words. Yet, experience shows that other words may be characteristic of the topic of interest. Discovery and use of those in text searches often leads to relevant documents. The authors propose a fully automatic method to discover such key-words from a set of texts deemed characteristic of the topic of interest. The method relies on n-gram coding of the texts, identification of characteristic n-grams in a subset of the texts, and finally searching for words containing one or many of the characteristic n-grams. An example using 6709 Reuters news briefs sampled from the 10 most common classes. Discriminating one class from the others serves to simulate the topic-of-interest situation. Each discrimination yields a set of "candidate key-words". This is compared to searching of topic-related words. Using n-grams appears to be more efficient as specific roots show up in the n-grams; it is automatic and does not require prior linguistic analysis.

1 Introduction

Documentary research (e.g. bibliography for a thesis or a research project technological monitoring, etc.) is often based on a few selected key-words. Yet, experience shows that other words may be characteristic of the topic of interest. Discovery and use of those in text searches often leads to relevant documents. [8] propose a method for text analysis based on several correspondence analyses interweaved with human interventions. In this paper, the authors propose a fully automatic statistical method to discover such key-words from a set of texts deemed characteristic of the topic of interest. The method relies on n-gram coding of the texts.

Several papers have shown the effectiveness of n-grams as a means of representing texts for clustering (partition in homogeneous groups) or categorization (assigning a text to one or many categories from a given list); see [10, 7, 3, 4].

Why are n-grams so efficient at classifying texts? How can one move up from shape to substance? This paper will help understanding why n-grams are efficient. The passage from the form to the meaning of a text is restored, moving from n-grams characteristic of a class of texts to the words comprising them.

Algorithm 1 Method for the selection of candidates key words

1. For each class j find all n-grams in all texts of the learning set,
2. Create the cross table of the N_{ij} occurrences of n-gram i in class j,
3. Compute the corresponding relative frequencies f_{ij}: $f_{ij} = \frac{N_{ij}}{N}$
4. Compute χ^2_{ij}, the contribution of cell (ij) to the χ^2 distance: $\chi^2_{ij} = \frac{\left(N_{ij} - \frac{N_{i.} \times N_{.j}}{N}\right)^2}{\frac{N_{i.} \times N_{.j}}{N}} = N \times \frac{(f_{ij} - (f_{i.} \times f_{.j}))^2}{f_{i.} \times f_{.j}}$
5. Compute $A_{ij} = \chi^2_{ij} \times sign(f_{ij} - (f_{i.} \times f_{.j}))$
6. Sort the A_{ij} by decreasing order
7. For each class j do
   (a) Create the list $\{gram_{lj}\}$ for $l = 1, ..., L$, of the L first n-grams of the class
   (b) for each $gram_{lj}$ do
      i. find all words ($words_{jk}$) such that $gram_{lj} \subseteq words_{jk}$
      ii. count the number $nb_{words_{jk}}$ of repetitions of $word_{jk}$ in the class
   (c) For each $word_{jk}$ do
      i. extract all $gram_{word_{jk}}$ included in $word_{jk}$, their total is noted $nbGram_{word_{jk}}$
      ii. For each gram $gram_{word_{jk}}$ do
      if $(gram_{word_{jk}} \in \{gram_{lj}\})$ then $presenceGram_{word_{worjk}}++$
   (d) if $\frac{presenceGram_{word_{jk}}}{nbGram_{word_{jk}}} > threshold_1$ and $nb_{words_{jk}} > threshold_2$ then $word_{jk} \in \{candidate\, key\, words\, for\, class\}$

The class	Acquisition	Earn	Money-fx	Wheat	Trade
Nb texts	1629	2841	528	209	362
The class	**Crude**	**Corn**	**Grain**	**Interest**	**Ship**
Nb texts	383	173	427	346	194

Table 1: Distribution of the number of texts among the 10 largest classes

The operational result is the automatic unveiling of a list of statistically characteristic words, that is, candidate key-words from which the user can select those to be retained. This work thus precedes that of [6] who, in pursuit of the same objective, had suggested an iterative method for the selection of words or of groups of words.

Steps needed in the detection of statistically characteristic words are described in section 2; two examples on rather large real life data sets follow; the final section indicates directions for future research. But first, let's review the principles of n-gram coding and its properties.

n-gram Coding A "n-gram" is a sequence of n consecutive characters. The set of n-grams (usually, n is set to 2, 3 or 4) that can be generated for a given document is basically the result of moving a window of n characters along the text. The window is moved one character at a time. Then, the number of occurrences of each n-gram is counted. For example, the phrase "The babysitter babysits the baby" can be represented by [the= 2, he_=2 , bab= 3, aby= 3, bys= 2, ysi=2 , sit=2 , itt= 1, tte= 1, ter= 1, er_= 1, r_b= 1, _ba= 3, its= 1, ts_= 1, s_t= 1, _th= 1, e_b= 2]. To simplify reading, the character "_" will be used to represent a blank.

Advantages of n-gram coding Techniques based on n-grams offer many advantages:

- Comparatively to other techniques, n-grams automatically capture the roots of the most frequent words [5]. There is no need for the identification of lexical roots (babysit, babysitter, babysitting, etc.).
- They work regardless of languages [4], contrary to word-based systems that require language-specific dictionaries (gender; singular-plural; conjugations; etc.). Moreover, n-grams do not require prior segmentation of the text into words; this is an interesting feature when dealing with languages where word limits are not clear, arabic language, for example.
- They are robust to spelling mistakes and distortions caused by optical text recognition. Scanned texts are often imperfect; for example, "chapters" could be recognized as "clapters". A word-based system might have some difficulty in recognizing "chapters" because of the erroneous spelling. Yet, a system using n-grams may still use the recognizable n-grams "apte", "pter", etc. [7] have shown that document search systems based on n-grams remained efficient in spite of a 30% distortion, a situation where no word-based system can operate correctly.
- Finally, n-gram techniques do not have to discard stop words and do not require stemming. These processes improve word-based systems. For n-gram based systems, studies have shown that they don't improve after stemming and discarding stop words [9].

Acquisition class		Crude class	
The most significant 3-grams	**Extracted key-words**	**The most significant 3-grams**	**Extracted key-words**
acq cqu qui uis iti sit	acquisition	oil _oi il_ il,	oil oil,
acq cqu qui uir	acquire acquired acquiring acquiring	rud cru ude	crude
sha har are	share	bar arr rel els	barrels
sha har are reh hol lde eho old der	shareholder holders holding hold	cua uad dor ado	ecuador ecuadorean
com omp any pan	companies company;	bpd _bp pd_ pd,	bpd (baril par jour)
tak ake eov keo	takeover stake take	gas	gas
sto toc ock lde	stockholders	ene erg rgy nergy_	energy
mer erg	merger; merge	pet etr leu eum ole	petroleum
off ffe fer	offer offers offering offered	plo xpl lor ora	exploration
has pur has	purchase	sau aud udi di_	saudi
usa sai air	USAir	zue ezu nez uel	venezuela
buy	buy buys	bbl	bbl (barrel)
inv nve sto	investment investment investor	pip ipe pel	pipeline
scl	disclosed undisclosed	xxo exx	exxon
cyc ycl	cyclops	ref ner efi	refinery
sac	transaction		
com omp ple let	complete	ara rab iea	arabian arabia
fil	filing	cub bic ubi	cubic
oup	group	_ku kuw uwa wai	kuwait kuwaiti
tst	outstanding	ric ice ces	prices
twa	twa (Trans World Airline)	ope pec	opec (non-opec)

Table 2: The first few significant grams and corresponding key-words for *Acquisition* and *Crude* classes

2 Identification of characteristic words

The key idea is to extract the n-grams typical to a class, then to retain the words containing the n-grams. The authors have written a Java program that looks for and counts n-grams among classes of texts, selects those that appear to be most typical of the classes, and then looks for words containing the typical n-grams and eliminates parasite words, see algorithm 1.

Identifying characteristic n-grams Before the complete algorithm is described, here are its principles. The key steps are: (i) identify all n-grams contained in all the texts of the learning set; (ii) build the cross table (text class×n-gram); (iii) compute (χ^2_{ij}) the cell contributions to the independence χ^2; (iv) for each class: identify characteristic n-grams (those significantly more frequent in some classes that in others); (v) search for words containing those n-grams.

While a number of statistics can be derived from the matrix of frequency (N_{ij}) of n-gram i in text class j, the χ^2_{ij} distance is often quoted as the most efficient in empirical comparisons [1, 11]. There are theoretical reasons for that; [2] showed that it is asymptotically equivalent to the "information gain".

In practice, the method suggested here yields a long list of words among which a number of parasite words, that is words, though otherwise uninteresting, happen to contain one of the characteristic n-grams. The objective, now, is to determine the list of "candidate key-words".

Filtering out parasite words In order to avoid "parasites", the process can be repeated backwards: for each word identified earlier, the n-grams it contains are examined and matched against the list of the n-grams characteristic for the class. If:

- the proportion of n-grams in the word matched to the list of n-grams characteristic of the class reaches some threshold, and
- the frequency of the word in the text also reaches some threshold,

then the word is considered to be a "candidate key word". If the word appears often on the text, it should be a candidate. If it occurs rarely and was selected because it contains one or two of the n-grams found in a common word, then it must be a parasite.

For example, a 3-gram like "*acq*" in the class of "Acquisition" will yield words characteristic of the class, like "acquisition" or "acquire"; but the 3-gram can also be found in uncharacteristic words, like "Jacques" or "racquets" that will be considered as parasites because they are rare in the class.

3 An Example

The proposed method should help the user to select documents of interest among an unstructured set of documents like the Web. First, the user must

The most significant 3-grams
[oil] [_oi] [bpd] [rud] [_bp] [il_] [cru] [pd_] [bar] [rel] [cua] [arr] [etr] [uad] [gas] [...] [rgy] [eum] [leu] [xpl] [els] [sau] [dor] [pet] [ado] [zue] [ezu] [0_b] [nez] [uel] [ira] [aud] [ole] [di_] [bbl] [ec_] [pip] [lor] [cks] [l_p] [pd,] [tpu] [plo] [utp] [gy_] [..1] [odu] [cub] [rod] [ude] [kuw] [uwa] [pd.] [n_b] [i_a] [_cr] [pel] [iea] [rre] [bic] [_ir] [ice] [xxo] [exx] [raq] [bia] [ner] [udi] [/bb] [ara] [ipe] [rab] [ene] [pd)] [mex] [obr] [thq] [hqu] [s/b] [uak] [_op] [duc] [al-] [ref] [(bp] [e_o] [ubi] [fie] [ait] [pec] [tro] [fue] [uot] [_ie] [ora] [ls_] [dri] [quo] [fsh] [efi] [eia] [mob] [as_] [exa] [pdv] [vsa] [dvs] [wai] [_ku] [_ga] [f_o] [ric] [um_] [_bb] [ces] [ukm] [dez] [iel] [urk] [aqi] [try] [xic] [uct] [abi] [naz] [rol] [xac] [a_b] [erg] [eik] [_dr] [put] [prt] [qi_] [c_m] [kh_] [ian] [tex] [ikh] [aeg] [ia_] [ia'] [_pd] [aq_] [rs/] [wti] [_km] [noc] [ope] [ubr] [il,]
Extracted key-words
[oil.] [oil,] [oil] [oilfield] [bpd] [(bpd)] [bpd.] [bpd,] [crude] [crude,] [crudes] [crude.] [oil prices] [oil industry] [oil companies] [oil prices,] [oil prices.] [oil price] [oil and] [oil production] [bpd in] [barrel.] [barrel] [barrels] [barrels.] [barrel,] [barrels,] [reliance] [ecuador,] [ecuador] [ecuador's] [ecuadorean] [petroleum] [petrobras] [petroleos] [petroleum,] [gasoline] [gas] [energy] [exploration] [exploratory] [exploration.] [saudi] [saudis] [venezuela] [venezuela,] [venezuelan] [venezuela's] [fuel] [iraqi] [iraq] [iranian] [iran] [iran's] [saudi arabia] [dlrs/bbl.] [opec] [non-opec] [pipeline] [pipeline,] [stocks] [output] [products] [product] [production] [production.] [producing] [producer] [produce] [producers] [produced] [products.] [products,] [production,] [cubic] [kuwait] [kuwait,] [kuwaiti] [mln barrels] [mln bpd] [iea] [current] [prices] [prices.] [price] [prices,] [exxon] [arabia] [arabian] [arabia's] [arabia,] [refinery] [general] [refineries] [refiners] [including] [arab] [mexico] [earthquake,] [earthquake] [operating] [opec's] [operations] [open] [reduction] [refining] [crude oil] [the oil] [because of] [price of] [fields] [field] [expected] [quota] [quoted] [barrels per] [barrels of] [barrels a] [drill] [drilling] [offshore] [mobil] [was] [has been] [as] [texas] [as a] [texaco] [pdvsa] [of oil] [american] [sources] [industry] [ministry] [a barrel.] [a barrel] [sheikh] [drop] [canadian]

Table 3: Complete list of class specific n-grams and candidate key-words for Crude class.

assemble a learning set, that is, provide two sub-sets of documents: one sub-set containing some documents of interest, and a second sub-set containing texts from the same source(s) but irrelevant to the task at hand.

As this work is user specific, the following example is general enough that the reader can easily understand and control, namely selecting news briefs on a given topic.

Reuters' indexed data The example uses Reuters press agency news briefs as a benchmark.

For the purpose of the example, the learning sub-set of 6709 news briefs from the 10 largest classes is derived from the "Apte version" comprising 7789 briefs [11]. Table 1 shows the size of each class.

Some results Results for Acquisition and Crude classes are displayed in Table 2. The method selected about 100 candidate key-words for each class; due to space constraint, only the first few significant grams and corresponding words are listed.

About the results using the Reuters collection For each class, the method proposes a list of class-specific candidate key-words, free of most parasites. The lists appear reasonable. Obviously, these results are in part due to the events of the time having generated the texts. The rule of "*all else being equal*" applies here as well.

The method is completely independent from the language of the text, since there is no need to remove spaces and punctuation marks.

No real improvement was noted from the different trials were conducted with either 1+2+3-grams or 4-grams did: results are quasi identical. This coincides with other authors' findings, for example [6, 3].

Two tables complete the exposition:

- The complete list of the class-specific n-grams for "crude" class are displayed in Table 3.
- A comparison of the result of four different text coding techniques using the proposed method is shown in Table 4:
  - Computation of χ^2_{ij} on the $(word_i \times class_j)$ cross table with elimination of spaces and punctuation,
  - Computation of χ^2_{ij} on the $(word_i \times class_j)$ cross table without elimination of spaces and punctuation,
  - Computation of χ^2_{ij} on the $(gram_i \times class_j)$ cross table with elimination of spaces and punctuation,
  - Computation of χ^2_{ij} on the $(gram_i \times class_j)$ cross table without elimination of spaces and punctuation.

It can be seen that the proposed method, based on n-grams without preprocessing, gives excellent results. It is hence completely independent from the language of the texts of interest.

4 Conclusion

The algorithm proposed in this paper is adept at extracting candidate keywords that are characteristic of a set of texts. An example is given on a set of 6709 Reuters news briefs organized in 10 classes (the 10 largest of the Reuters collection). The proposed method gives good results as it selects keywords sharing characteristic n-grams. A method to remove parasite words, based on the proportion of significant n-grams they contain, is also described. The method is efficient although it operates on raw texts, without any prior linguistic analysis.

text coding techniques	**Extracted key-words**
Extracted key-words **from complete words with** punctuations and spaces elimination	[oil] [bpd] [crude] [opec] [barrels] [barrel] [ecuador] [energy] [exploration] [petroleum] [prices] [gasoline] [gas] [refinery] [saudi] [saudis] [pipeline] [production]
Extracted key-words **from complete words without** elimination of the punctuations and spaces	[oil.] [oil,] [oil] [crude] [opec] [opec's] [non-opec] [barrels] [barrels.] [barrels,] [bpd] [(bpd)] [bpd.] [bpd,] [energy] [petroleum] [ecuador,] [ecuador] [ecuador's] [exploration] [gasoline] [gas] [refinery] [saudi] [saudis] [prices] [prices.] [prices,] [barrel.] [barrel] [barrel,] [cubic] [production] [production,] [output] [stocks] [drilling] [pipeline] [pipeline,] [today] [day] [days] [yesterday] [iea] [arabia] [arabian] [natural] [venezuela] [venezuelan] [texaco] [petrobras] [api] [herrington] [mobil] [exxon] [offshore] [iranian] [feet] [15.8] [quota] [refining] [reserves] [kuwait] [wells] [fuel] [fields] [industry] [field] [iraqi] [minister] [spot] [demand] [price] [lukman] [santos] [producing] [iraq] [shell] [sources] [texas] [rigs] [research] [sea] [iran] [greece] [gulf]
Extracted key-words from **n-grams with** elimination of the punctuations and spaces	[oil] [bpd] [bp] [crude] [crudes] [oil prices] [oil industry] [oil stocks] [oil companies] [oil minister] [oil company] [oil price] [oil and] [oil production] [bpd in] [barrel] [barrels] [ecuador] [ecuadorean] [petroleum] [petrobras] [petroleos] [petro-canada] [gasoline] [gas] [energy] [exploration] [exploratory] [levels] [saudi] [saudis] [venezuela] [venezuelan] [fuel] [iraq] [iranian] [iran] [000 barrels] [000 bpd] [saudi arabia] [bbl] [pipeline] [stocks] [output] [products] [product] [production] [producing] [producer] [produce] [producers] [produced] [cubic] [kuwait] [kuwaiti] [iea] [mln barrels] [mln bpd] [current]
Extracted key-words **from n-grams without** elimination of the punctuations and spaces	[oil.] [oil,] [oil] [oilfield] [bpd] [(bpd)] [bpd.] [bpd,] [crude] [crude,] [crudes] [crude.] [oil prices] [oil industry] [oil companies] [oil prices,] [oil prices.] [oil price] [oil and] [oil production] [bpd in] [barrel.] [barrel] [barrels] [barrels.] [barrel,] [barrels,] [reliance] [ecuador,] [ecuador] [ecuador's] [ecuadorean] [petroleum] [petrobras] [petroleos] [petroleum,] [gasoline] [gas] [energy] [exploration] [exploratory] [exploration.] [saudi] [saudis] [venezuela] [venezuela,] [venezuelan] [venezuela's] [fuel] [iraqi] [iraq] [iranian] [iran] [iran's] [saudi arabia] [dlrs/bbl.] [opec] [non-opec] [pipeline] [pipeline,] [stocks] [output] [products] [product] [production] [production.] [producing] [producer] [produce] [producers] [produced] [products.] [products,] [production,] [cubic] [kuwait] [kuwait,] [kuwaiti] [mln barrels] [mln bpd] [iea] [current] [prices] [prices.] [price] [prices,] [exxon] [arabia] [arabian] [arabia's] [arabia,] ...

Table 4: Comparisons of four techniques on the class "crude".

References

[1] K. Aas and L. Eikvil. Text categorization: a survey. Technical report, Norwegian Computing Center, 1999.

[2] J. P. Benzecri. *L'Analyse des Données*, volume 2. Dunod, Paris, 1976.

[3] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US, 1994.

[4] T. Dunning. Statistical Identification of Languages. Technical Report MCCS 94-273, Computing Research Laboratory, 1994.

[5] Gregory Grefenstette. Comparing Two Language Identification Schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data (JADT'95)*, Rome, Italy, 1995.

[6] A. Lelu and M. Hallab. Consultation "floue" de grandes listes de formes lexicales simples et composées : un outil préparatoire pour l'analyse de grands corpus textuels. In M. Rajmann and J. C. Chappelier, editors, *JADT'2000*, volume 1, pages 317–324, Lausanne, mars 2000.

[7] E. Miller, D. Shen, J. Liu, and C.Nicholas. Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System. *Journal of Digital Information*, 1(5), 1999.

[8] Annie Morin. Intensive use of correspondance analysis for information retrieval. In *ITI'04*. To appear.

[9] Mehran Sahami. *Using Machine Learning to Improve Information Access*. PhD thesis, Computer Science Department, Stanford University, 1999.

[10] Olivier Teytaud and Radwan Jalam. Kernel based text categorization. In *Proceeding of IJCNN-01, 12th International Joint Conference on Neural Networks*, Washington, US, 2001. IEEE Computer Society Press, Los Alamitos, US.

[11] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90, 1999.

Address: ERIC Laboratory, Lyon 2 University
5, av. Pierre Mendès-France F-69676 Bron - FRANCE
E-mail: {rjalam, chauchat}@univ-lyon2.fr; jean.dumais@statcan.ca