Textual data mining by parsing

A. Bellacicco

Department of Systems, Theory and Organization, Teramo University, Italy.

Abstract

The paper deals with the problem of the identification of the specific answer to whatever question put to a search engine in the web exploration. The actual problem is to avoid the usual stream of thousands pages extracted by the search engine without the help of an a priori categorization of the theme to which the question is directed.

The system, called IRAS, described in the paper, is a follow up of a previous theoretical framework, called TOM, presented at MIS 2002, Bellacicco [1].

The innovation of IRAS is the exploitation of parsing algorithms for the identification of the semantic organization of the statements so that the answer is specific for the real content of the question.

Problem faced by IRAS is therefore to overcome besides the stream of thousands pages from the web, the stream of unspecific answers too, which are the fallout of the ambiguity related the use of the terms of the query without the specification of their role in the statement. The parsing of the query besides the parsing, selection and clustering of the statements is the primary tools of IRAS.

1 Introduction

Data mining and textual mining differ mainly by the use of tools which consider numerical operations besides the logical operations and by the use of tools linked to operators which consider the specific role of the data.

For the numerical case the operators are functors which map the data either into a geometric hyperspace or on an hypesurface. The choice of the geometry
depends from the outside informations on the problem so that data are weighted, compressed or distributed according to the available informations.

2 The geometry of textual mining

Textual data are characterized by operators which map each statement into a combinatorial frame which is a subframe of a frame at an higher level. The higher level is still a combinatorial frame. The geometry of the mentioned combinatorial structures can be envisaged as a group theoretical frame where the groups are actually permutation groups and the operators maps groups of permutations of symbols into clusters of ordered sequences of terms. To each cluster a specific set of labels can be attached which can be considered their meaning.

The logic of search is therefore based on the identification of homomorphisms among subsets of statements and the query which respect the permutation group structure of both of them.

The clusters of answers are therefore viewed as homomorphic sequences of symbols and the query behaves as a template permutation group which draws the homomorphic images from the pages extracted by a search engines.

We can consider the previous considerations as a way of thinking the selection of the statements which can be the answer to a query as a particular case of cryptography both on the terms and on the statements, Dass and Eugeni [2].

The query behaves like a junction of two public keys at both the levels so that the selection becomes first the search of the statements which are the candidates to the coding by the query treated as a statement and second the selection of the best one among the candidate statements as the optimally coded statements.

This strategy actually runs as a sequence of two steps of optimal coding considering all the available statements provided by a search engine as a restricted set of possible decoders of the query. The geometry of the codes are an abstract way of thinking beyond the pure combinatorial one, Beutelspacher and Eugeni [3].

IRAS can be considered therefore as a decoding system restricted to a finite set of possible solutions.

We remark here that the symbols used both for text, that is for instance, the query and for the codes are sequences of terms among the terms available from all the selected statements and for the statements the set of all allowed statements which are characterized by the presence or the absence of specific modifiers. The role of the modifiers can be conceived as a filter which select the available statements selected by the search engine.

The parsing becomes now a powering tool both of the selection of the statements and of the right answer.
We will describe the parsing process following the algebraic point of view outlined in the previous pages.

3 Coding and decoding texts by parsing

Texts can be considered as a set of statements which are finite permutations of subsets of terms. Actually each statement is characterized by a few subset of terms whose role depends on the position in the statement. Depending on the chosen language we have small particles, that is terms which qualify the role of the subsequent term. Such particles may precede or may follow a given term and may be joined to a given term either before or after.

The grammatical parsing has the target of identifying the inflection of a common noun, the identification of the adjectives and of the verbs and the conjugation of a verb so that we can distinguish their temporal role. Regarding the inflection, we must recall here that the sequence of two or more common nouns is qualified by their inflection so that their ordering is not enough for distinguishing their meaning.

Apart from the special case of proper nouns, common nouns denote sets so that the inflection means set theoretical operations on sets.

i. The genitive case means that a common noun $P$ is considered as a predicate of a set represented by another common noun $Q$ which refer to a generic set $X$.

ii. The dative case means a mapping $f$ from $X$ to a set $Y$.

iii. The accusative case means $X \rightarrow Y$, where $\rightarrow$ means imply.

iv. The ablative case means the inverse mapping $f^{-1}$ into $Y$ from $X$.

v. The nominative case identifies the primary set $X$.

The verb are the operators $\rightarrow$ and $f$ and code a specific action.

Following the point of view of considering the query as a template, that is a public key in the optimal coding, the strategy outlined for TOM and now considered as a strategic tool for IRAS, the permutation group whose kernel is the query, is characterized by two segments of terms, where the first segment admit a fixed correspondent while the second segment admits substitution through a set of permutations.

Following the previous identification of the various cases of a declination of the nouns and the consequent inflection as set theoretical operations we can easily transform the set operations into logical operations so that we can write formulas in the frame of first order predicate calculus.

Something should be said about the universal and existential quantifiers and overall on the negation which can admit various interpretations. We recall that all the following asserts are true where $\neg$ means not:

i. $\neg \forall x P = \exists x \neg P$

ii. $\neg \exists x P = \forall x \neg P$

iii. $\forall x P = \neg \exists x \neg P$

iv. $\exists x P = \forall x \neg P$. 
The negation is the consequence of a philosophical choice on the exclusion principle. The choice implies the truth of the four asserts and moreover the exclusion of vague asserts and probabilistic asserts. The inclusion both of probability and of vagueness implies the inclusion of the Bayes theorem and its inversion formula. Bayes theorem implies stochastic inference in the narrow sense.

The inclusion of the measure of probability modifies the way we use the well formed formulas in first order logic as far as we must suppose for each formula a mapping on the real interval which can be the unitary closed interval $I$ in the probabilistic case and the non negative real axis in the vagueness evaluation.

In any case both the probabilistic evaluation and the vagueness evaluation belongs to the metalanguage which is the natural language and their inclusion in the fist order logic implies the inclusion in a finite frame of the infinite frame of the real numbers. Moreover the mixture between the natural language which is the target of the logicism whose aim is its transformation into logical formulas and the same natural language used for evaluating the truth of asserts and so on implies a *regressio ad infinitum* introducing a paradox in the strategy of coding. This inclusion does not allow therefore the coding/decoding strategy outlined in the previous pages. The previous asserts allow the use of just one quantifier, that is either $\forall$ or $\exists$. Moreover, the operators $\rightarrow$ and $f$ can be reduced to the operations of inclusion and exclusion.

Something should be said about verbs and articles. A verb can be considered as an operator which can assume the meaning either of implication or of mapping. The implication too can be viewed as a mapping and the difference between the two cases is given by the image of the mapping which is a subset in the former case and a whole set in the latter case.

Usually the type of verb implies the use of the dative case or the use of the accusative case. The small particles *at, to, on, onto* are customized in the actual use of a verb so that the distinction between genitive and dative case depends from the current use of a verb.

In the same way the articles can distinguish the existential operator from the universal operator and therefore can be viewed as the quantifiers whose role are well defined by the four asserts mentioned earlier.

Some conclusions can be drawn:

i. a well formed formula in the first order calculus of predicates is able to translate quite all the statements in the natural language;

ii. grammar and syntax of a western language, like english, german or spanish can be parsed so that all the terms roles can be translated into a logical frame;

iii. the strategy used for identifying the right answer to a question put to a search engine can be considered as a coding/decoding strategy provided
the comparison between the query and the candidate statements is treated as a permutation group on a finite set of basic symbols which are at the highest level the available propositions of the opened pages and at a lower level the parsed terms.

iv. Parsing allows the satisfaction of the exact meaning as far as the permutation group is constrained by the grammatical rules so that some sequences of terms are constrained and the set of possible permutations is restricted to a subset which satisfies the grammatical rules.

4 Coding and decoding

First problem to analyse is the computational complexity of the parsing and coding process. IRAS actually treats the result of the primary search of a search engine and its input is the set of all the statements rescued by a search engine and into the statements all the used terms.

The main problem is the identification of the leading particles which qualifies the type of query and its object. Terms like what, when, who, where, how much, how old,... and few other terms identify the type of request to which must correspond respectively

- an object: the statement must refer to a physical entity;
- a date: the statement must contain a temporal reference of an event;
- a noun: the statement must refer to a living organism like a man, a woman and so on;
- a place: the statement must contain the physical location of an event;
- a numerical evaluation of a predicate P.

Inside all the rescued statements containing at least a term or a couple of terms written in the advanced search language; using the logical operators without a previous knowledge of the theme of search by the parsing process, all the terms and their grammatical and syntactical role, are considered like the brick of a wall of different shape. The combinatorial analysis of the strings of terms belonging to each statement. IRAS is able to identify the statement which fits the query first selecting the statements which map the type of query and secondly through permutation of the basic terms suitably parsed so that the constraints exclude all the permutation which does not fit the basic permutation group generated by the query.

We can turn back to cryptography. The basic approach to cryptography still today is transposition and substitution when the primary stones to deal with are the same both for the original text and for the coding tools. Usually for an alphabetic text the classical approach is the use of the same alphabet following a different rule which use the alphabetic items for substituting still other alphabetic items following basically the old ideas of the roman emperors and back in the ancient Egypt.
The strategy of IRAS is to search automatically among the statements the statement or the statements which can be considered the answer to the query.

The query is viewed as the public key for decoding the rescued statements and choosing the best one through a permutation process which fit the query to each statement so that the query decodes at the best the statement so that the identified statement can be considered the right answer to the query, Mc Eliece[4].

The asymmetry comes out from the chance of reversing the process so that each statement is a key and the strategy is to find the best key. Following this point of view the asymmetric asset of the treatment of the keys allows the use of the query for decoding the rescued pages or vice versa it is possible to use each selected statement which satisfies the required correspondence in the leading terms for decoding the query.

From the computational point of view there is no difference between the two approaches as far as the number of required operations depends from the selected number of statements and from the number of terms witch at most is equal to the number of available terms belonging to a dictionary of the chosen language. In the real practice the number of terms is equal to the number of all the terms belonging to the rescued pages. Following this point of view we must consider that the decoding operations use the available statements and the best fit is given by the occurrence of quite all the terms occurring in the query so that a primary filter is given by a statistical evaluation of the pairing of the highest number of common occurrences.

Some observations can come out considering the semantic alternatives so that many terms not occurring can have the same meaning. We follow here a strategy used for coding the written mapping of a spoken language. In this case the classes of equivalence of the occurring terms can enlarge the decoding process.

From the combinatorial point of view, we consider a permutation group as the direct product of all the permutation groups associated to the available terms. This strategy can be enriched considering all the terms belonging to a permutation group as key words for detecting automatically the pages not yet rescued.

This strategy can be considered cumbersome unless we select a subset of permutations reducing the permutation classes only to the key words used for activating a search engine. In this case the number of permutations is definitively low and is not required a feed-back to the search engine.

From the operational point of view the recommended strategy consider the query as the key so that all the possible offsprings generated by a key word used for writing down the query can be considered a new interpretation that is a semantic interpretation of the key words.

Considering for each key word $w_i$ a permutation class of size $n_i$ and if the key words are $k$, we get the set of possible permutations:

$$(\prod n_i!)!$$
In case $k = 3$, $n_i = 3$ we get $6^3 \cdot 3! = 216 \cdot 3!$. We consider the basic permutations group generated by the query and we consider only 216 copies of the permutation group of the query. In our case the number of operations drastically reduces to 1296 which is absolutely low.

We have to compare 1296 templates with the set of available statements which can be for instance $10^2$. The number of comparisons grows up to 129,600 which is still a feasible number.

The previous evaluations come out from the selection of the statements filtered by the leading particles which characterise a statement and the query. The filtering of the available statements becomes a strategic choice whose impact on the number of operations is of great deal as far as it reduces drastically the search.

5 How IRAS improves TOM

IRAS actually improves quite all components of TOM both from a theoretical point of view and from the computational point of view, Bellacicco [1], as far as the conception of IRAS as a decoder with a public key of a text considered as a cryptic message.

In order to see how this assertion is true we can see the components following their role in a session of IRAS. The graph of the logical layout of the functional organisation of IRAS is shown in figure 1 where the functional roles of each component is evidenced.

The main improvement is due to the introduction of a decoding action based on the availability of grammar rules enriched by lemmas and of syntax rules.

The parsing process on the key words of the query and the strategy of comparison with the statements rescued by a search engine is reduced in IRAS into a combinatorial sequence of attempts of matching. Ferraro and other [5].

The choice of the best solution is just a cluster of solutions whose leading seed is the set of key words and the set of recurrences of the key words or their interpretations exploiting the intrinsic ambiguity of each term in a language L.

We mentioned that the polyhedric meaning of each term is reduced to a class of homonymous terms which belong to the rescued statements.

The parsing process builds up besides the grammar and the lemmas of each term the class of the alternative meanings of each term and the homonymous terms for each meaning of the term.

All the mentioned tools are lodged in the knowledge base of IRAS so that the iteration of the decoding process can be fast and really competitive with a common user.
IRAS LOGICAL LAYOUT

Figure 1: Layout of the IRAS components and operations.
The iteration implies the choice between the decoder and the answer identifier and clustering.

References


