Text and Web Mining Approaches in Order to Build Specialized Ontologies
Mathieu Roche, Yves Kodratoff

To cite this version:

HAL Id: lirmm-00424463
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00424463
Submitted on 15 Oct 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Text and Web Mining Approaches in Order to Build Specialized Ontologies

Mathieu Roche\textsuperscript{a,*}, Yves Kodratoff\textsuperscript{b}

\textsuperscript{a}LIRMM, CNRS, Univ. Montpellier 2, 34392 Montpellier Cedex 5 – France
\textsuperscript{b}LRI, CNRS, Univ. Paris-Sud, 91405 Orsay Cedex – France

Abstract

This paper presents a text-mining approach in order to extract candidate terms from a corpus. The relevant candidates are selected using a web-mining approach. The terms (i.e. relevant candidate terms) we find are the instances of specialized ontologies built during this process. The experiments are based on real data – Human Resources corpus – and they show the quality of our text and web mining approaches.

Key words: Text-Mining, Web-Mining, Terminology, Statistical Measures, Conceptual Classification

1. Introduction

Our approach extracts knowledge directly from the initial corpus. It is made up of two main phases (1) collecting a homogeneous corpus on a given topic, (2) building a categorization specific to the corpus. The first phase will not be described in this paper, since it is not performed in a completely automated way. It is important however not to forget that this phase is the basis upon which the whole text-mining process is built, since its success depends heavily on the quality and the homogeneity of the collected corpora. The various treatments described in this paper have been carried out on a French corpus on the topics of Human Resources (3,784 KB). This corpus is provided by the PerformanSe company.\textsuperscript{1}

The second phase of the text-mining process is based on the identification of terms in the texts. The search for terms finds the significant word groupings for the specialty field. The terms obtained are associated to concepts, i.e., those showing the same semantics are clustered. Each cluster represents a concept, meaningful to the field expert, and these basic concepts can be, and usually are, the first level of an ontology of concepts. The existence of more general concepts is asserted by the expert, when needed. For instance, our expert recognized from the Human Resources corpus that the set of French terms \textit{vérifiable solidarité} (genuine solidarity), \textit{travail en groupe} (group work), \textit{réseau de relations} (network of relationships) were the linguistic observables indicating the existence of the concept Relational in the texts. These indications of presence of a concept are defined as "instances" of this concept. We thus obtain an ontology, the nodes of which are concepts.

In order to build the specialized ontology, we apply a text and web mining process to extract the
instances of the concepts. The text-mining method
evaluations to extract candidate terms (section 2) and the web-
mining approach validates the extracted candidates (section 3).

2. A text-mining approach in order to

extract terminology

2.1. Cleaning and PoS tagging

In text-mining approaches all the corpora we worked
upon had in common to be largely unsuitable to further
linguistic treatment due to the variety of unexpected forms
they contain. For instance in the Human Resources corpus,
the vocabulary has been normalized. The main point to signal
is that the writing of cleaning procedures cannot be done
without the help of field expert who provides sets of
rules specific to the domain. For instance, in scientific
domains (e.g. Biomedical domain), the complex
combinations of upper and lower case letters convey
meaning the expert only can deal with.

The PoS (Part of Speech) tags each word of
the cleaned texts with a grammatical label. During this
step, we used Brill’s tagger [2]. After tagging, we
are able to extract the doublets or triplets of single
words, showing a specific grammatical label (Noun,
Adjective, Preposition, etc).

2.2. Extracting terms

This step automatically extracts the terms from
the texts (i.e. relevant candidate terms). In our
work we have used the Exit system to extract terminology
[10]. This system enables us to extract
the candidate terms Noun-Noun, Noun-Adjective,
Adjective-Noun, Noun-Preposition-Noun from a

2 http://www.lri.fr/~heitz/formulaire_logiciels.html

3 Test of English as a Foreign Language
To obtain scores, PMI-IR uses several measures based on the proportion of documents where both terms are present. Turney’s formula [15] is inspired by Mutual Information described in section 3.2.2. The measure calculates the proportion of documents containing both word and choice_1 (within a 10 words window), and compares with the number of documents containing the word choice_2. The higher this proportion is, the more word and choice_1 are seen as synonyms. This kind of web technique will be used in our work. But as argued in the next section, we use several quality measures to order the candidates.

3.2. Statistical Measures

Several quality measures in the literature are based on ranking functions. They are brought out of various fields: Association rules detection [1,8], terminology extraction [5,9], and so forth. The following are the most widely used.

3.2.1. Number of Occurrences

The basic measure FR used is based on the number of occurrences nb(x, y). This function corresponds to the number of web pages provided by the query "x y" with a search engine (in our work, we use the Exalead\(^4\) search engine). Actually, the value returned by this function corresponds to the popularity of the use of the two words x and y together like a string. Note that the Turney’s approach calculates the dependency of the words using the AND and NEAR operators of the search engine.

3.2.2. Mutual Information

One of the most commonly used measures to compute a kind of relationship between the words composing what is called a co-occurrence is Church’s Mutual Information (MI). The formula is the following [3]:

\[
I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}
\]  

(1)

Such a measure tends to extract rare and specific co-occurrences according to [5,9,14]. Let us notice that in this formula (1), the use of the \(\log_2\) function is not mandatory, since the latter is strictly growing. Thus, the order of the co-occurrences provided by the measure is not impacted by the application of the \(\log_2\) function. In our context, \(P(x, y)\) measures the probability of finding couples of words \((x, y)\) where \(x\) et \(y\) are neighbors, and in this order. When simplified, the formula (1) could be written as follows, where \(nb\) designates the number of occurrences of words and couples of words:

\[
MI(x, y) = \frac{nb(x, y)}{nb(x)nb(y)}
\]  

(2)

3.2.3. Cubic Mutual Information

The Cubic Mutual Information is an empirical measure based on MI, which enhances the impact of frequent co-occurrences, something which is absent in the original MI [4]. Such as measure is defined by the following formula:

\[
MI^3(x, y) = \frac{nb(x, y)^3}{nb(x)nb(y)}
\]  

(3)

Vivaldi et al. have estimated that the Cubic MI was the best behaving measure [16]. This measure gave good results in our work regarding the desambigisation of the acronym definitions [12].

3.2.4. Dice’s Coefficient

An interesting quality measure is Dice’s coefficient [13]. It is defined by the following formula:

\[
D(x, y) = \frac{2 \times P(x, y)}{P(x) + P(y)}
\]  

(4)

Similarly to the Cubic MI, Dice’s coefficient weakens the impact of rare and often irrelevant co-occurrences [11]. Formula (5) leads directly to formula (4)\(^5\), which relies on the \(nb\) occurrences of words and couple of word occurrences:

\[
Dice(x, y) = \frac{2 \times nb(x, y)}{nb(x) + nb(y)}
\]  

(5)

3.3. Using a context

In this paper, we define the context as a set of significant words given by the expert to define the field. These words are added to the queries using the AND operator in the statistical measures. For instance, without the context, the \(nb(x, y)\) function calculates the number of pages provided by the search engine by the query "x y". With the context based on the

\[\frac{nb(x, y)}{nb_total}\]

http://www.exalead.fr/search/
words \(c_1, \ldots, c_n\), we apply the query "\(x \text{ AND } c_1 \text{ AND } \ldots \text{ AND } c_n\)". The goal of this context is to restrict the searching space to the Web pages of the domain.

4. Experiments

4.1. An ontology used as a benchmark

We have extracted terminology regarding the specialized Human Resources corpus. We can select the most frequent terminology: Candidates that appear only one or two times are not taken into account, i.e., they are pruned [11]. The Table 1 shows the results of extracted candidates in our corpus.

<table>
<thead>
<tr>
<th>Patterns</th>
<th>number of occurrences</th>
<th>number of occurrences after pruning at 3</th>
<th>rate of pruning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun-Prep-Noun</td>
<td>4,703</td>
<td>1,268</td>
<td>73.0%</td>
</tr>
<tr>
<td>Noun-Noun</td>
<td>98</td>
<td>11</td>
<td>88.8%</td>
</tr>
<tr>
<td>Adjective-Noun</td>
<td>1,290</td>
<td>476</td>
<td>63.1%</td>
</tr>
<tr>
<td>Noun-Adjective</td>
<td>5,768</td>
<td>1,628</td>
<td>71.8%</td>
</tr>
</tbody>
</table>

Table 1 Results of occurrence number of extracted candidate terms.

The expert manually analyzes the candidates extracted by the EXIT system in order to build a specialized ontology. Then our Human Resources ontology is composed of 3,000 terms (instances of the concepts of the first level of the built ontology). Our ontology has 3 levels: 12 concepts for the first level, 6 7 concepts for the level 2, and 2 concepts of the level 3. The following figure shows a representation of our ontology:

The terms (i.e. instances of the concepts) manually validated will be used as benchmark in order to evaluate our web-mining approaches.

4.2. Results of the web-mining approaches

We have studied the 500 most frequently extracted candidates in our Human Resources corpus. 73% of these candidates were evaluated as relevant by an expert. Note that the frequent candidates in a corpus are often relevant [11].

We have evaluated the rate of relevant candidates (precision) after applying statistical measures presented in section 3.2. In Table 2, the quality measures are evaluated according to different thresholds (i.e. precision calculated with the \(n\) first candidates given by statistical measures). This Table shows that all the statistical measures give better results than a random classification (e.g., the precision is increased by 9% up to 15% for the first 100 words).

Table 3 shows the sum of the ranks of relevant candidates using the different measures. When the value is low, the results is better. Note that minimizing the sum of the ranks of relevant candidates is

---

6 French concepts: environnement, relationnel, Vous-Même, communication, stress, rôle, indépendance, influence, hiérarchie, Comportement&Attitude, activité, activité_gestion&administration
equivalent to maximizing the Area Under the ROC curves (AUC - Area Under the Curve). This criterion of AUC [6] is often used to evaluate learning algorithms [7] or ranking functions [9]. We have calculated the sum of the ranks with and without the use of a context, the French word "psychologie" (i.e. "psychology") in our queries. This Table enables us to observe two important results:

– The use of the context improves the result for all measurements.
– Dice’s measure gives better results with and without context.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Random measure</th>
<th>FR measure</th>
<th>MI measure</th>
<th>MI$^4$ measure</th>
<th>Dice’s measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.73</td>
<td>0.86</td>
<td>0.88</td>
<td>0.92</td>
<td>0.88</td>
</tr>
<tr>
<td>100</td>
<td>0.73</td>
<td>0.82</td>
<td>0.88</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td>150</td>
<td>0.73</td>
<td>0.77</td>
<td>0.80</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>200</td>
<td>0.73</td>
<td>0.78</td>
<td>0.78</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>250</td>
<td>0.73</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>300</td>
<td>0.73</td>
<td>0.77</td>
<td>0.79</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>350</td>
<td>0.73</td>
<td>0.78</td>
<td>0.77</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>400</td>
<td>0.73</td>
<td>0.77</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>450</td>
<td>0.73</td>
<td>0.75</td>
<td>0.74</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>500</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2
Precision of web measures taking into account the French context "psychologie".

<table>
<thead>
<tr>
<th>Measures</th>
<th>FR</th>
<th>MI</th>
<th>MI$^4$</th>
<th>Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td>without context</td>
<td>87,584</td>
<td>88,789</td>
<td>87,502</td>
<td>87,066</td>
</tr>
<tr>
<td>with context</td>
<td>87,146</td>
<td>87,105</td>
<td>86,815</td>
<td>86,597</td>
</tr>
</tbody>
</table>

Table 3
Sum of the ranks of the relevant candidates.

5. Conclusion

The approach presented here extracts a terminology using a text-mining approach. The second step of our process selects relevant candidates using a web mining technique based on statistical measures. The results show that the association of statistical measures with a context improves the results. Furthermore, Dice’s measure seems well suited (e.g., 80% of candidate terms are relevant based on the first 200 candidates returned by this measure). These relevant candidate terms represent the instances of the built ontology. In our future work, we wish to take into account other statistical measures. We will develop a web-mining approach to automatically determine the context. This richer context could improve the results obtained with statistical measures.

Acknowledgments

We thank Serge Baquedano (PerformanSe company) for the acquisition and the expertise of the Human Resources corpus.

References


