Flexible Document Categorisation

JEBAIRI CHAKER
Department of Computer Science
College of Computer and Information Sciences
King Saud University
P.O. BOX 51178 Riyadh 11543
KINGDOM OF SAUDI ARABIA

OUNALLI HABIB
Département d’Informatique
Faculté des Sciences de Tunis
Université de Tunis El’Manar
Campus Universitaire 2092 Tunis
TUNISIE

Abstract: - In the context of automatic document categorisation, we propose in this paper a new flexible approach for electronic document categorisation situated in junction of knowledge engineering and learning machine approaches. Our approach assigns a HTML document to one or more categories (paper, call for papers, email, …) using three types of criterions: physical, logical and discursival criterions. Using a set of pre-categorised document, this approach generates a base of categorisation rules. This base is used to categorise new documents. The categorisation flexibility is carried out with rule weight association representing your importance in the discrimination between possible categories. This weight is calculated using the Zadeh min t-norm and it’s dynamically modified at each new categorisation. The proposed approach is experimented using a corpus of 615 HTML documents belonging to different predefined categories. The obtained results are satisfactory and make up a primary validation for our approach.

Keywords: - Document, category, learning, discrimination, flexible, physical, logical, discursival, rule.

1 Introduction

In front of the incredible growth of web documents, we notice that following a search query, the actual information retrieval systems provide a large list of documents. In front of this list, the user cannot quickly select relevant documents. Automatic document categorization is a fundamental component in many information systems, because he allows the organization of documents by category. Several automatic document categorization methods have been proposed in the literature. These methods can be divided in two approaches: knowledge engineering and machine learning methods.

In this paper, we propose a new flexible categorization approach. This approach assigns a French HTML document to one or more predefined categories (paper, call for papers, email, …).

Our proposed approach can respond many motivations:
- Accelerate information retrieval systems.
- Improve thematic classification accuracy.
- Facilitate the assimilation and dissemination of great information loads.
- Allows the application of suitable summarization method.

In this paper we start with presenting a brief survey of related works in the automatic document categorization. In the third section, we present the principle of our approach. In the fourth, fifth and sixth sections, we explain respectively the three principles steps of the proposed approach. The approach experimentation is also presented in the seventh section. Finally, we propose in the conclusion some future works.

2 Related works

The automated document categorization dating back to 60 years, with Maron works [13]. Since then, several authors have proposed different categorization concept definitions [1, 11, 12, 18, 19]. According to Sebastiani [18, 19], the categorization of documents set D consists in assigning each document d belonging to D a category c belonging to a set of predefined categories C.

Automatic document categorization has been used in a number of different applications: automatic indexing for Boolean information retrieval systems, document organization, word sense disambiguation, yahoo-style search categorization [18].
Several automatic document categorization methods have been proposed in the literature. These methods can be divided into two approaches: knowledge engineering and machine learning methods. Maron has proposed knowledge engineering approach in 1961 [13]. It is based on categorization rules of type IF Condition THAN Category [2, 9]. This approach has been abandoned because it needs a manual effort to build and manage the set of categorization rules. To solve this problem the categorization community have proposed in 1980 to use some machine learning techniques [14]. The principle of this last approach consists in automatically generating a categorization function using a set of training documents. This function is used to categorize new documents. Among machine learning algorithms we mention: Rocchio’s algorithm [8], K-Nearest Neighbor [5], Naïve Bayes [1, 8], Decision trees [3, 16], Support Vector Machines [10, 20], Voted classification [4, 7].

3 Principle of proposed approach

Our proposed approach assigns a French HTML document to one or more predefined categories (dictionary, patent, book, thesis, memory, report, paper, FAQ, call for papers, web page, news, email) using three types of criterions: physical (the document size in number of words), logical (the document logical structure) and discursival (the existence of some linguistic expressions in some predefined logical units). Our approach is situated in junction of the knowledge engineering and machine learning approaches. Using a set of training documents, our approach allows automatically generating a categorization function. This function is represented in the form of a base of categorization rules. Contrary to other methods such as Decision trees [3, 15, 16], Galois lattice [6] or induction graphs [17, 22], where graph transformation in rules is necessary. In our approach we have three types of rules: physical, logical and discursival rules exploiting respectively physical, logical and discursival criterions. Each categorisation rule is of type IF Condition THAN Conclusion. The categorization flexibility is carried out with rule weight association representing your importance in the discrimination between possible categories. This weight is calculated using the Zadeh min t-norm [21] and it's dynamically modified at each new categorization. The approach principle is illustrated in figure 1 below.

4 Generation of categorisation rules

4.1 Training collection

To generate categorization rules, we collect from web a training set A of 1230 French HTML documents. Each training document $d_j$ is represented by: the identification $did_j$, the category $C_j$, the document size in number of words $nm_j$, the logical structure $sl_j$ and the discursival unit $ud_j$. The distribution of the training set $A$ on the 12 possible categories is presented in table 1 below.
Table 1. Number of training documents by category

<table>
<thead>
<tr>
<th>Notation</th>
<th>Category</th>
<th># Of training documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Dictionary</td>
<td>30</td>
</tr>
<tr>
<td>C2</td>
<td>Book</td>
<td>40</td>
</tr>
<tr>
<td>C3</td>
<td>Patent</td>
<td>40</td>
</tr>
<tr>
<td>C4</td>
<td>Thesis</td>
<td>100</td>
</tr>
<tr>
<td>C5</td>
<td>Memory</td>
<td>100</td>
</tr>
<tr>
<td>C6</td>
<td>Report</td>
<td>100</td>
</tr>
<tr>
<td>C7</td>
<td>Paper</td>
<td>120</td>
</tr>
<tr>
<td>C8</td>
<td>FAQ</td>
<td>100</td>
</tr>
<tr>
<td>C9</td>
<td>Call for papers</td>
<td>100</td>
</tr>
<tr>
<td>C10</td>
<td>News</td>
<td>160</td>
</tr>
<tr>
<td>C11</td>
<td>Web page</td>
<td>180</td>
</tr>
<tr>
<td>C12</td>
<td>E-mail</td>
<td>160</td>
</tr>
</tbody>
</table>

4. 2 Categorisation criterions

4. 2.1 NM criterion
NM criterion represents the document size in number of words. Using the training set A, we have identify three symbolic values of this criterion:
High if NM > 50000; Medium if NM ∈ [5000,50000] and Low if NM < 5000

4. 2.2 SL criterion
Logical units ordered one after the other to appear an idea. For each logical unit we have associated a weight between 0 and 1 representing your importance in the logical structure construction. This weight is calculated using the training documents. For this criterion we have identified 9 possible logical structures (see table 2 for some examples).

Table 2. Predefined logical structures

<table>
<thead>
<tr>
<th>SL</th>
<th>Content</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Titre (1), auteur(s) (1), affiliation(s) (1), email(s) (1), résumé (1), mot clés (1), introduction (1), texte (1), conclusion (1), remerciements (0.2), références (1).</td>
<td>Paper</td>
</tr>
<tr>
<td>S2</td>
<td>Titre (1), date et lieu (1), introduction (1), thèmes abordés (1), soumission (1), comité scientifique (1), comité d’organisation (1), dates importantes (1), informations (0.8).</td>
<td>Call for papers</td>
</tr>
<tr>
<td>S3</td>
<td>Destination (1), sujet (0.8), texte (1).</td>
<td>e-mail</td>
</tr>
</tbody>
</table>

4. 2.3 UD criterion
A discursival unit is a set of linguistic expressions which is present in a predefined logical unit. A discursival unit is verified if at least one of their linguistic expressions is present in at least one logical unit. Using training documents we have identified 12 possible discursival units (see table 3 for some examples).

Table 3. Predefined discursival units

<table>
<thead>
<tr>
<th>UD</th>
<th>Linguistic expressions</th>
<th>Logical units</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ud1</td>
<td>Cet article, le présent article, l’article suivant, ce papier, Le présent papier, le papier suivant.</td>
<td>Résumé+introduction</td>
<td>Paper</td>
</tr>
<tr>
<td>Ud2</td>
<td>Cet AàC, l’AàC suivant, le présent AàC, cet AàP, c’est AàP, l’AàP suivant, le présent AàP.</td>
<td>Titre+introduction+thèmes abordés</td>
<td>Call for papers</td>
</tr>
<tr>
<td>Ud3</td>
<td>Ce message, cet e-mail.</td>
<td>texte</td>
<td>e-mail</td>
</tr>
</tbody>
</table>

4. 3 Physical rules
Physical rules generation is based on NM criterion values. In our approach we have 3 physical rules. Each rule is of type:

IF Condition(NM) 

THAN \{(C_1, \alpha_1), (C_2, \alpha_2), \ldots, (C_{12}, \alpha_{12})\}

Where: Condition(NM) is verified if NM ∈ {high, medium, low} and \(\alpha_i\) is the appartenance degree to the category \(C_i\). This degree is the proportion of training documents, which belongs to the category \(C_i\) and verifies Condition (NM).
4. 4 Logical rules
Using the logical criterion values, we have identified 9 possible logical rules. Each rule is of type:

IF Condition(SL)
THAN \{ (C_1, \beta_1), (C_2, \beta_2), \ldots, (C_{12}, \beta_{12}) \}

Where: Condition(SL) is verified if \text{Argmax}_{SL} \ \text{SIM}(sl_j, sl_i) \geq S_0 \text{ and } \beta_i \text{ is the appartenance degree to the category } C_i. \text{ This degree is the proportion of training documents, which belongs to the category } C_i \text{ and verifies Condition (SL).}

SIM(sl_j, sl_i) \text{ is the similarity between document logical structure } sl_j \text{ and the predefined logical structure } sl_i. \text{ This similarity is calculated using this formula:}

\text{SIM}(sl_j, sl_i) = \sum_{i : \text{ul}_i \in sl_j} p_i \sum_{i : \text{ul}_i \in sl_i}

Where: \text{ul}_i \text{ is a logical unit belonging to the predefined logical structure } sl_i, \text{ and } p_i \text{ is the weight assigned to the logical unit } \text{ul}_i \text{ and } S_0 \text{ it’s the similarity threshold, under this value Condition(SL) is not verified. In our case, we have chose a threshold value as 0.5.}

4. 5 Discursival rules
Using the discursival criterion values, we have identified 12 possible discursival rules. Each rule is of type:

IF Condition(UD)
THAN \{ (C_1, \gamma_1), (C_2, \gamma_2), \ldots, (C_{12}, \gamma_{12}) \}

Where: Condition(UD) is verified if UD \in \{ ud_1, \ldots, ud_{12} \} \text{ and } \gamma_i \text{ the appartenance degree to the category } C_i. \text{ This degree is the proportion of training documents, which belongs to the category } C_i \text{ and verifies Condition (UD).}

5 Categorization of new documents
At each new document \text{d}_j, \text{ the categorization process identify the document size in number of words } \text{nm}_j, \text{ your logical structure } sl_j \text{ using } <Hn> \text{ tags and verify the existence of some predefined discursival units } ud_j. \text{ After this preprocessing, three discrimination types may be executed: physical, logical and discursival exploiting respectively physical, logical and discursival rules.}

The physical discrimination allows the use of the physical rule that verifies Condition(nm_j). The result of this discrimination is a first possible list of categories:

\text{C}_1 = \{ (C_1, \alpha_1), (C_2, \alpha_2), \ldots, (C_{12}, \alpha_{12}) \}

In second step, the logical discrimination allows the generation of second possible list of categories, using a suitable logical rule which verify Condition(sl_j):

\text{C}_2 = \{ (C_1, \beta_1), (C_2, \beta_2), \ldots, (C_{12}, \beta_{12}) \}

Finally, the discursival discrimination allows the generation of a third list of possible categories, using the suitable discursival rule which verify Condition(ud_j):

\text{C}_3 = \{ (C_1, \gamma_1), (C_2, \gamma_2), \ldots, (C_{12}, \gamma_{12}) \}

To obtain the optimal categorization \text{C}^*, \text{ we should combine the three sets } \text{C}_1, \text{ C}_2 \text{ and } \text{C}_3. \text{ This combination is carried out using the Zadeh min t-norm.}

\text{C}^* = \text{C}_1 \cap \text{C}_2 \cap \text{C}_3 = \{ (C_1, \text{min}(\alpha_1, \beta_1, \gamma_1)), (C_2, \text{min}(\alpha_2, \beta_2, \gamma_2)), \ldots, (C_{12}, \text{min}(\alpha_{12}, \beta_{12}, \gamma_{12})) \}

6 Modification of categorization rules
After each new categorization, we should update the set of rules. This modification is summarized in two fundamental points, which are:

- Remove rules, which their conclusions are equal to 0. In other words, the rules whose all their appartenance degrees to all possible categories are equal to 0.
- Since, the proportion of training documents verifying Conclusion rules will be modified. We should recalculate the appartenance degrees for all rules.

7 Experimentation
Our proposed approach has been implemented in the CFD system. To experiment this system, we have used a corpus of 615 HTML
documents belonging to the possible categories. For each testing document d_j. We have identified the three criterions nm_j, sl_j, ud_j. Exploiting these criterions, we have obtained these results (see table 4).

Table 4. Category Recall, Precision, Accuracy and Error

<table>
<thead>
<tr>
<th>Category</th>
<th># Of testing documents</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Category Recall</th>
<th>Category Precision</th>
<th>Category Accuracy</th>
<th>Category Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0.7</td>
<td>0.87</td>
<td>0.67</td>
<td>0.33</td>
</tr>
<tr>
<td>Book</td>
<td>10</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.8</td>
<td>0.89</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>Patent</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0.6</td>
<td>0.86</td>
<td>0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>Thesis</td>
<td>10</td>
<td>27</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0.9</td>
<td>0.93</td>
<td>0.84</td>
<td>0.16</td>
</tr>
<tr>
<td>Memory</td>
<td>33</td>
<td>33</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.94</td>
<td>0.97</td>
<td>0.92</td>
<td>0.08</td>
</tr>
<tr>
<td>Report</td>
<td>50</td>
<td>48</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0.96</td>
<td>1.00</td>
<td>0.96</td>
<td>0.04</td>
</tr>
<tr>
<td>Paper</td>
<td>70</td>
<td>70</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>FAQ</td>
<td>10</td>
<td>10</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0.94</td>
<td>0.96</td>
<td>0.91</td>
<td>0.09</td>
</tr>
<tr>
<td>CIP</td>
<td>60</td>
<td>55</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0.92</td>
<td>0.98</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>News</td>
<td>100</td>
<td>90</td>
<td>5</td>
<td>10</td>
<td>4</td>
<td>0.9</td>
<td>0.95</td>
<td>0.86</td>
<td>0.14</td>
</tr>
<tr>
<td>Web page</td>
<td>90</td>
<td>70</td>
<td>5</td>
<td>20</td>
<td>5</td>
<td>0.78</td>
<td>0.93</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>Email</td>
<td>80</td>
<td>77</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0.96</td>
<td>0.94</td>
<td>0.91</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Where:
A: The number of documents correctly assigned to the predefined category.
B: The number of documents incorrectly assigned to the predefined category.
C: The number of documents incorrectly rejected from the predefined category.
D: The number of documents correctly rejected from the predefined category.

We have obtained a recall average value of 0.87, a precision average value of 0.94, an accuracy average value of 0.84 and an error average value of 0.16. These results are satisfactory.

9 References


8 Conclusion and future works

We observed that the use of some metadata criterions can improve categorization accuracy of French HTML documents. In the future works, we propose:

- The integration of new electronic formats (SGML XML, …) to exploit the meta data provided by the Dublin Core norm.
- The integration of this approach in the process of information retrieval to improve their performance.


