Automated Detection & Identification of Textual Areas in Dematerialization Process

Mohammed MOUJABBIR
Department of computer System
Faculty of sciences and technology
Mohammedia, Morocco
moujabbir@gmail.com

Mohammed RAMDANI
Department of computer System
Faculty of sciences and technology
Mohammedia, Morocco
ramdani@fstm.ac.ma

Abstract

Documents Management System (DMS) are systems that read information depending on document’s categories and suggest item. While the majority of existing researches compute recommendation by considering several methods such as: classifiers, learning methods, clustering methods etc…

We conduct experiments to identify the best detection algorithm for our dataset. We evaluate our approach from 10 types of documents. We focus on recognition the entire element of textual sectors –items contained in area text- . Results shows that the algorithm based on the calcul of similarity provide an acceptable rate of detection area. We conclude that this research area detection needs deeper studies.

Keywords: Documents dematerialization, electronic invoicing dematerialization, character automatic recognition, Ontology, Similarity concept
1. Introduction

In order to provide the efficiency of DMS especially in relevant of the results, future generation of this kind of system should integrate multitudes functionalities: intelligence, adaptability, expressivity, indexing etc...

Document dematerialization processes are dedicated to the improvement of the efficiency, inexpensiveness and accessibility of all users’ document: dematerialization activities introduced to manage bureaucratic digital documents in a proper way, are among the main tasks of the e-government, e-commerce, accounting works.

In this paper we present a novel method in the dematerialization process, that constitutes the detecting all the sector area content by using the similarity concept able to identify correctly the most existing area in the digital document. The simulation results show that the proposed algorithm can effectively reduce the ranking abnormality and the number of handoffs.

This paper is organized as follows. Section II presents a description of related work. Section III describes the pre-processing and document representation phase. Section IV presents our experimental design based on similarity approach. Section V includes the discussions and critical. Section VI concludes this paper.

2. Related work

2.1 Relevant terms recognition

The state of the art in this domain is related to techniques of NLP (Naturel Langage Processing) including several approaches:

- Statistical Linguistics[7][11].
- Computational Linguistics [6][17][22][30].

Whose objective is the study and the analysis of natural language and its functioning through computational tools and models. In particular, for the analysis of the textual areas, however specific disciplines have been developed, like Corpora Linguistics and Textual and Lexical Statistics [1][12].

Term extraction presents the first stage in automatic document processing and derivation of knowledge from texts. The second stage is the analysis of the lexical items which carries a specific meaning.

Generally Terms serve to convey the fundamental concepts of a specific knowledge domain for example the football term refers to the sport concept: they use their relationships to build the semantic meaning.

Methods for term extraction from texts can be divided in three main categories:

- Linguistic methods: They use the linguistic knowledge, heuristic rules for identification tokenisation, and harmonisation the tokens. They use also a part-of-speech tagging, word-stemming, lemmatization and identification of phrase-structures [1][12];
Statistical methods: They analyse the word occurrences within texts in order to measure the “weight” of the selected term. The termhood[21] allows the assignment of weight, it represent the degree by which a term is related to the specific concept of the domain. Unithood measures the collocation strength of the units; Evaluating unithood is based on measuring significance of the words occurring [38][10].

Hybrid methods: They use two methods: linguistic and statistical, the first is based on part-of-speech tags, to extract a set of candidate terms, the second assigns a value to each candidate term[23].

2.2 Segmentation
The segmentation can be performed by means of special instruments; it can be formed by two fundamental components:
- Glossaries listing well-known expressions to be regarded as tokens;
- Grammars containing heuristic rules (in the form of regular rules), which are manually written by experts.

The combined use of glossaries and grammars ensure high levels of precision. However results depend on the kind of the document, textual content and language used: texts which are full of acronyms or abbreviations can increase the mistakes rate. Consequently, the glossary and the grammatical rules should be adapted to the characteristics of the domain at issue.

In the literature, several works focus on relevant knowledge. In contrast various systems deal the extraction knowledge by different means and use various methods. The used methods are described in the next section.

2.3 Systems based on the distributional properties of words
They consist in analyzing the distributions of words in order to recover the semantic distance between the concepts represented by those words. The distance can be used as a:
- Hierarchical clustering to automatically derive hierarchies of concepts form texts [14][25].
- Formal Concept Analysis[9].
- Classification of words inside existing ontologies [4][28].
- Learn concept hierarchies [8] [36] [26] [3].

2.4 Systems based on pattern extraction and matching
They rely on lexico-syntactic patterns to discover semantic relations between words in unrestricted texts. Hearst [18] pioneered using patterns to extract hypernymy relations; Berland and Charniak [5] applied the previous technique to extract meronymy. Girju [16] have studied meronymic relations extraction while Turney [35] has proposed a uniform approach for the extraction of different kinds of relations from text.

Hearst [19] proposes to look for co-occurrences of word pairs appearing in a specific relation inside WordNet. Turney[34] presents an unsupervised learning algorithm that mines large text corpora for patterns expressing implicit semantic relations.
2.5 Hybrid approaches

They combine statistical and pattern-based techniques, as in Alfonseca and Manandhar [4], Ryu and Choi [29] have proposed an algorithm for IS-A relation extraction from the English Wikipedia. Giovannetti [15] propose a methodology that integrates lexico-syntactic patterns, manually defined, (pattern-based approach).

2.6 Ontologys

Ontology is a formal explicit specification of a shared conceptualization of a domain of interest [33].

An Ontology aims at providing a formal and explicit description of the concepts in a domain of discourse. The main elements of the ontology are concepts, relationships and instances:
- Concepts represent the categories and the classes of objects that are relevant in the domain of interest;
- Relationships serve to semantically connect concepts and instances;
- Instances represent the named and identifiable concrete objects in the domain of interest (i.e. the particular individuals that are classified by concepts and related by relationships).

There exist several approaches for ontology types classification, proposed amongst others by Lassila & McGuinness [24] in 2001 and by Oberle [27] in 2006. In this paper we characterize the ontology using 3 dimension:
- Number of conceptual elements in the domain.
- Degree of ambiguity in the conceptualization of the domain.
- Expressiveness of the formalism used to specify the ontology.

As we can clearly see, none of the existing techniques meets all our requirements. Therefore there is a need for a novel approach. In fact existing approaches do not allow: Document processing invoice format, Processing of documents in case of error injection

We are inspired by these methods to develop a new method based on the calculation of similarity. This calculation takes several aspects into consideration: syntactic similarity, the Synsets similarity and the hierarchical similarity.

3. Preprocessing and document representation

After collection of a given data set, this one has to be processed before we could perform identification and detection tasks on its documents. This process consists of two main steps. The first step focuses on pre-processing techniques. The second step concerns the identification and the detection the entire area existing in the document.

3.1 Preprocessing

This phase consists of applying a number of techniques so as to clean and normalize the raw textual documents. We describe below, briefly, the techniques of pre-processing stage.
Automated detection & identification of textual areas

- **Tokenization:** Basically, the main objective is to transform a text document to a sequence of tokens separated by white spaces. The output of this process is a text free from punctuation marks and special characters. The most common model is the bag-of-words which consists on splitting a given text into single words. We can also find the n-gram word model[13]. As a definition, an n-gram word is a sequence of n words in a given document. If n is equal to 1, we talk about unigrams, if it is equal to 2, they are called bigrams, and so on. For example, in the sentence “Facture N° 35/2013”, the unigrams that we can extract are “Facture”, “N°”, “35”, “/” and “2013”. While the bigrams that we can find are “Facture”, “N°”, and “35/2013”.

- **Stemming:** Word stemming is a crude specific language process which removes suffices, synonym, sign in order to reduce words to their word stem[2]. For example, the stemming of the two words “Num” and “N°” will give the same stem “Numéro”.

- **Stop Words elimination:** Stop Words are function words which provide structure in language rather than content. For example The list of stop words can be “. # ^ etc...”.

### 3.2 Document representation

After pre-processing stage, we proceed to adaptive format building. This textual format consists of all terms that occur after all pre-processed documents. These terms can be words and/or expressions. The following figure shows an example of document representation.

![Figure 1: document representation](image)

### 4. The proposed method

As a pre-processing step, we have removed from words all special characters, like: punctuations marks, separators, spaces etc... This feature will not contribute in identification process. This is why we can say that these features do not have discriminate power and it is recommended to remove them.

In this study, we are interested in a detection and identification of textual area in the given document, where each of them is characterised by several features as: position, morphology, information relevance, vulnerability and sensibility to recognition errors.

For our tasks of area’s identification, we use the data from ABBYYFineReader [32], it is one of the strongest and widely optical character recognition (OCR)
used in the dematerialization of documents. We test three different classifiers based on:

- Jaccard’s Coefficient [20].
- Grammatical Distance [31].
- Hierarchical similarity [37].

4.1 Critical

The use of each variable independently cannot in any case give a satisfactory decision, because the problem addressed: automatically identification of area—depends on several constraints:

- Syntactic constraints: it specifies the degree of syntactic similarity, ignoring the semantic goal.
- Morphological constraints: it specifies the degree of similarity in sub-tree of the used ontology, also called the degree of correlation.
- Constrains related to the Synsets: it specifies the degree of similarity between the terms entered in the ontological tree and the area to detect.
- For those reasons we propose in the next section a formula to evaluate the score of each area, it will be used in the proposed algorithm.

4.2 Hierarchical similarity

It represents the quotient of the number the nodes in the Z model on the number the nodes in the area $Z_i$. The used nodes belong to the considered ontology.

Formally:

$$SimHear(Z, Z_i) = \frac{NumberOfNodes(Z_i)}{NumberOfNodes(Z)}$$
Example:

We consider the following model area:

\( Z_{\text{Designation}} \) = the designation model area and

**Number Of Nodes**\( (Z_i) = 6 \)

Let’s evaluate the following areas text:

- \( Z_1 = \text{“casa to rabat 5 2000.00 10 000.00”} \)  
  \( \text{SimArch}(Z_{\text{Designation}}, Z_1) = 6/6 = 1 \)
  
  Casa ⇔ token11 ; to ⇔ token12 ; rabat ⇔ token13 ; 5 ⇔ token2 ; 2000.00 ⇔ token3 ; 10 000.00 ⇔ token4;

- \( Z_2 = \text{“rabat to 5 2000.00 10 000.00”} \)  
  \( \text{SimArch}(Z_{\text{Designation}}, Z_2) = 5/6 = 0.8333 \)

- \( Z_3 = \text{“casa Le 10/02/2013”} \)  
  \( \text{SimArch}(Z_{Designation}, Z_3) = 7/6 = 1.16 \) so
  \( \text{SimArch}(Z_{Designation}, Z_3) = 0 \) i.e. we will not consider this value.

**4.3 Synsets similarity**

This similarity is centered on the calculation of the Jaccard’s coefficient among the subset of synsets of the model area and \( Z \) subset of Synsets the zone \( Z_i \).

Formally:

\[ \text{SimSyn}(Z, Z_i) = Jaccard(Z, Z_i) = \frac{\text{card}(Z \cap Z_i)}{\text{card}(Z \cup Z_i)} \]
4.4 Syntactic similarity

It is based on the syntactic stack, where a syntax error (i.e. an empty case in the analysis table or no conformity in two tokens), then we subtracted from the syntactic similarity the \(-1/n\) value.

4.4.1 Reminders

In order to define the language denoted by a grammar symbol we need to define the set of alphabet then the concept of derivation. Derivation is a relation that holds between two forms, each a sequence of grammar symbols.

Formally, a grammar is a set denoted as \(G=(V_TV_N,S,P)\).

\- \(V_N\) is a non-empty set of terminal symbols.
\- \(V_T\) is a non-empty set of non-terminal symbols with \(V_N \cap V_T = \emptyset\)
\- \(S\) is an initial symbol (axiom).
\- \(P\) is a set of production rules

Example
\(V_T=\{a,c\} V_N=\{S,T\} S \rightarrow a T b \mid c \quad T \rightarrow c S S | S\)

Let’s calculate the set of firsts and set of Nexts

Definitions:

\- **first** \((A)\): the set of all terminal symbols that can start a chain that is derived from \(A\)

\- **Next** \((A)\): the set of all terminals that can appear immediately to the right of \(A\), \(S\) symbol belongs to the **Next’s** of the axiom. The following table shows the set of firsts and set of Nexts

<table>
<thead>
<tr>
<th></th>
<th>Firsts</th>
<th>Nexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>(a,c)</td>
<td>(S)</td>
</tr>
<tr>
<td>T</td>
<td>(a,c)</td>
<td>(b)</td>
</tr>
</tbody>
</table>

Table 1: The firsts and the nexts set

- **Analysis table** : is a two-dimensional array \(M\) which indicates for each non-terminal symbol \(A\) and a terminal symbol to the rule of production applied

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(S)</td>
<td>(S \rightarrow a T b)</td>
<td>(S \rightarrow c)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(T)</td>
<td>(T \rightarrow S)</td>
<td>(T \rightarrow c S)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: The analysis table
4.4.2 Example
We consider the following grammar:
\[ S \rightarrow \text{id} \text{id} \text{Ref} \ (1) \]
\[ \text{Ref} \rightarrow \text{chiffre} \text{Sep} \text{chiffre} \ (2) \]
\[ V_N = \{ S, \text{Ref} \} \ V_T = \{ \text{id}, \text{chiffre}, \text{Sep} \} \]
Then we generate the following tables:

<table>
<thead>
<tr>
<th>Firsts</th>
<th>Nexts</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{S}</td>
<td>\text{id}</td>
</tr>
<tr>
<td>\text{Ref}</td>
<td>\text{Chiffre}</td>
</tr>
</tbody>
</table>

Table 3: The firsts and the nexts set

<table>
<thead>
<tr>
<th>id</th>
<th>chiffre</th>
<th>Sep</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{S}</td>
<td>(1)</td>
<td></td>
</tr>
<tr>
<td>\text{Ref}</td>
<td>(2)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: The analyse table

There is four empty cases i.e. four syntactic errors possible
- Let's analyse the following sentence:
  \[ Z_1 = "\text{Facture N° 30/2013}" \]

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{S}</td>
<td>id id chiffre sep chiffre$</td>
<td>\text{S} \rightarrow \text{id id Ref}</td>
</tr>
<tr>
<td>$\text{Ref id id}</td>
<td>id id chiffre sep chiffre $</td>
<td>\text{Ref} \rightarrow \text{chiffre Sep chiffre}</td>
</tr>
</tbody>
</table>

Table 5: the analyse process

So the \text{SimSynta}\(Z_{\text{NumFacture}, Z_1}=1\)
- Let's analyse the following sentence:
  \[ Z_2 = "\text{Facture 30 N°/2013}" \]

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{S}</td>
<td>id chiffre id sep chiffre$</td>
<td>\text{S} \rightarrow \text{id id Ref}</td>
</tr>
<tr>
<td>$\text{Ref id id}</td>
<td>id chiffre id sep chiffre $</td>
<td>Syntax error: we need to find an id instead of a number</td>
</tr>
<tr>
<td>$\text{Ref}</td>
<td>id sep chiffre $</td>
<td>Syntax error: we need to find an number instead of a id</td>
</tr>
<tr>
<td>$\text{chiffre seq chiffre}</td>
<td>chiffre seq chiffre $</td>
<td>The sentence is not accepted</td>
</tr>
</tbody>
</table>

Table 6: the analyse process

So the \text{SimSynta}\(Z_{\text{Zone NumFacture}, Z_2}=1-(1/4)-(1/4)=1/2=0,5\)

In the first case, there is no error, but in the second case there are two errors.

4.5 Area comparaison

In order to detect all area contended in the document, we have considered several grammatical rules. These rules contain various features: lexical, syntactic and semantic. To achieve this task we specify for each category of document a set of rules called grammar. The similarity or association stuck between the sentences is centred by means of:
• Jaccard’s coefficient.
• Grammatical distance.
• Hierarchical similarity.

Jaccard’s coefficient is utilized to measure the intersection of two sets: grammatical rules and candidate sentence.

We make labeling areas as identification problem. For the given area $Z$, the problem is to find the model (also the class) of that area, that is to say the attribute to which it corresponds. The identifier process is based on the calculation of the degree of similarity. Similarity that takes into account multiple constraints: hierarchical similarity, syntactic similarity and lexical similarity.

Formally we use:

$$Sim(Z, Z_i) = \alpha \times SimHeiar(Z, Z_i) + \beta \times SimSynta(Z, Z_i) + \gamma \times simSynset(Z, Z_i)$$

Where $Z_i$ is the textual area to defined.

Therefore the score $Z_i$ will be defined like:

$$Z_{i0} = \text{ArgMax}_{Z_j \in Z_{modele}} \left( Sim(Z_j, Z_i) \right)$$

$Z_{i0}$ means the area getting the greatest value of $Sim(Z_j, Z_i)$

Where:

• $\alpha$, $\beta$ and $\gamma$: Designate a fixed constants belonging to the interval $[0.1]$ and $\alpha + \beta + \gamma = 1$
• $Z_{modele}$: Designate the set of model area
• $SimHeiar(Z_j, Z_i)$: Refers to the similarity of number of nodes in the ontological tree.
• $SimSynta(Z_j, Z_i)$: Designate the syntactic similarity, i.e. the degree of conformity the $Z$ and $Z_i$.
• $simSynset(Z_j, Z_i)$: Designate the similarity of common Synsets.

5. Experiments

In this section, we present the experiments that we have performed for this study. Firstly, we describe the experiments environment, i.e. some details about the pre-processing techniques, the identifiers, the validation method and the performance measure that we have used. Secondly, we present the obtained results and we discuss the different findings.

5.1 Process of experimentation

We scanned 200 documents from 10 categories, and we used ABBYYFineReader to find the text format–fig 3-. The considered categories are:
Automated detection & identification of textual areas

Transports invoices, wood invoices, irons invoices, invoices of building materials, invoices of spare parts, fuel invoices, invoices tires, invoices of notaries, lawyers invoices and phone invoices

We note in this sample, the existence of several errors of recognition, syntactic, lexical and semantic errors.

Figure 3: the text format

5.2 Algorithm of detection

In order to detect the type of all areas constituting the text document, we use the score calculation with the following algorithm:

1. For a zone $Z$ gives calculate the score ($Z$)
2. If the score is satisfactory
   a. the area is properly recognized
   b. move to the next area
3. Otherwise $Z$ together with another area go to 2

5.3 Results

We consider the following symbols that mean the model existing area:

$Z_1$=date area.
$Z_2$= number of invoice number.
$Z_3$= customer address area.
$Z_4$= designation label area.
$Z_5$= designation item area.
$Z_6$= amount ht area.
$Z_7$= amount ttc area.
$Z_8$= amount in words area.

The following table shows the results found, including the rate of detection zones.
The identification result that we obtained by the use of similarity method is 78.05%. The use of the multiple coefficient of similarity gives a good rate of detection and identification the area text that contained all useful data. The detection areas allows the recognition of all the components of this area, and also the possibility to obtain the best interpretation.

### 6. Conclusion and future works

In this paper, we investigate the use of similarity techniques to detect the contextual text area from document (invoices document) in order to provide ameliorations and recommendations to the developers of the DMS. We propose detection and classification approaches centred on similarity concept.

We evaluated our approach on our own dataset from multiple provider invoices. We focused on the areas text that we aim to detect the entire token contained, and they will be using as labels information in the interpretation phase.

We consider that the obtained result on our corpus, though it is not high, is promising given the challenges encountered in this area and that we mentioned above. Our result proves also that this research area especially the detection area text is challenging and still needs deeper and more fine-grained studies.

### References


Mohammed MOUJABBIR and Mohammed RAMDANI

545. Nantes, France.


[25] Lee L. (1997), Similarity-Based Approaches to Natural Language Processing, Ph.D. thesis;


[31] Stavros Papadopoulos Introduction to Data Mining Fall 2011


Received: June 19, 2013