Toward a Complex System for Context Discovery to Index Arabic Documents

Mohamed Salim El Bazzi^{1*}, Driss Mammass¹, Abdelatif Ennaji², Taher Zaki¹

¹ Laboratory IRF-SIC, University Ibn Zohr, Agadir, Morocco.

² Laboratory LITIS, University of Rouen, Rouen, France.

* Corresponding author. Tel.: (+212)653838855; email: elbazzi.mohamedsalim@edu.uiz.ac.ma

Manuscript submitted October 31, 2017; accepted January 20, 2018.

doi: 10.17706/jcp.13.8.955-962

Abstract: Text indexing aims to take the full advantage of textual data to help intelligent programs to make relevant decisions. In order to explore a large amount of textual documents, and to disclose semantic information hidden in unstructured documents, like texts, an effective indexation system is required. In this paper, we propose a new approach for indexing Arabic texts. Based on the semantic proximity and taking into account the contexts contained in each document, our method is denoted contextual indexing. Several algorithms are used for keywords extraction, each of them emphasizes some criterion. However, we target the most descriptive keywords for each document. We also propose a new approach for document modeling. We compared the results obtained using our method with those obtained by an indexation system based on a standard statistical method. The experimental results demonstrate the performance of our approach.

Key words: Contextual indexation, semantic proximity, clustering, Arabic documents.

1. Introduction

Indexing an amount of documents is to select its most representative descriptors in order to generate the list of indexing terms. It is a way of retrieving all the terms characterizing a document. Document indexing is an important step in the text mining process, because it determines how the knowledge contained in the documents is represented [1]. It takes place each time a document is added to the corpus.

The amount of information increases and the access to relevant information become a complex task. Clearly, the problem is not about the availability of information but its relevance to a particular context of use.

The context is the linguistic entourage of a textual element within the utterance in which it appears, that is to say, the series of units preceding it and following it. The term context refers to all the circumstances in which an act of enunciation takes place: cultural and psychological situation, experiences and knowledge of the world, Mutual representations that each one makes of his or her interlocutors, etc.

In this paper, we present a new approach based on context discovery in order to select the most relevant keyword that describe best a given document. We denote this approach "Contextual indexation".

The rest of this paper is organized as following. The second part introduces the related works. The third part is dedicated to the presentation of the Contextualization System. We will present the experimental results in the fourth section and finally we conclude by discussing our contribution and any possible improvements.

2. Related Works

Relevance is undoubtedly the fundamental question posed in text mining tasks. This subjective notion, which is essentially dependent on the user's point of view, has again been the subject of investigations within text classes in general and contexts in a particular way.

We can note a lack of representation of the context in text mining models especially for unstructured data. Classic models have been defined to assign a document to an already known classe.

The context is not a new concept in computer science, language theory and artificial intelligence already exploit this notion. With the emergence of information retrieval systems, the term is rediscovered and placed at the heart of debates without, however, being the subject of a clear and definitive consensus definition.

Hence, we define the context as a set of elements (phoneme, morpheme, sentence, etc.) which precede and / or follow a linguistic unit within a utterance.

The method we propose in this paper is mainly inspired by the nature of appropriate keywords, namely understandable semantically and relevant to the document.

Mesleh *et al.* [2] applied an ANT colony optimization (ACO) as a feature space reduction mechanism with χ^2 as a score function and then classified the Arabic documents using the SVM classifier.

Thabtah *et al.* [3] set up an Arabic categorization system using the naive Bayesian classifier based on the weighting characteristics provided by the χ^2 test to classify a simple labeled database. The experimental results, compared to the classified data set, show that the selection of characteristics often improves the accuracy of the classification by removing empty or rare terms.

Bawaneh *et al.* [4] compared the two classifiers, KNN (K Nearest Neighbor) and NB (Naïve Bayesian). The light stemmer was used as a characteristic and the TFIDF measurement as a weighting method for the characteristics. The KNN classifier was judged to be more efficient.

Kanaan *et al.* [5] classified the documents in Arabic with the expectation-maximization (EM) algorithm. The TFIDF measure is applied as a method of weighting the characteristic elements, while the naive Bayesian algorithm is used to calculate the labels of the documents, and finally the classification is made using the EM algorithm.

Raheel *et al.* [6] combined the Boosting method and the decision tree as a hybrid classifier. They used lemmatisation as a method of extracting the characteristics, and the TFIDF for the weighting. A comparison of the method was made with two classifiers, Bayesian Naïve (NB) and SVM (Support Vector Machine). The result shows that SVM and NB surpass the proposed approach.

Zaki *et al.* [7] extend the vector space model by combining the TF-IDF with the Okapi formula to extract relevant concepts that represent a document. It proposes a new measure that takes into account the notion of semantic neighborhood by using a measure of similarity between terms and combining the calculation of the TF-IDF-okapis with a core approach (radial-based function). This indexing approach allows a contextual and semantic search.

Al-Salemi *et al.* [8] used characteristics selection techniques such as mutual information, statistics χ^2 , information gain, ESG coefficient And Odds Ratio to reduce the size of feature space by eliminating items that are considered irrelevant for a category being studied.

Mansour *et al.* In their work [9], perform a morphological analysis of the document to extract the indexes. The authors propose a process of extraction of stems on the one hand, and of the nouns and verbs on the basis of rhymes and grammatical rules. A weight is then assigned to each stem taking into account its occurrence and introducing a function indicating how the word is spread out in the document.

Tomokyo and Hurst [10] propose a method that verifies grammaticality (a key word must be well-formed syntactically) and informativity (KL divergence. Thus, for a candidate term, the greater its

probability of passing from the uni-gram model generated from the analyzed document to the N-gram. Similarly, the greater its probability of passing from the N-gram model from a reference corpus to an N-gram model generated by the processed document, the more the term candidate is informative.

Jamoussi [11] proposes a method of extracting key words based on the semantic representation of terms. In her work, Jamoussi presents two methods based on semantic distances, the Kullback-Leibler distance and the average mutual information to calculate the amount of information between two words or two classes of words. The new method introduced by Jamoussi is tested against a simple vector representation, with three unsupervised classifiers: the K-means algorithm, Kohonen maps and the Bayesian AutoClass network

As for zaki *et al.* [12], they introduce in there work the notion of semantic neighborhood. It proposes a hybrid system for the contextual and semantic indexing of Arabic documents, bringing an improvement to the classical models based on n-grams and the Okapi model. It calculates the similarity between the words using a hybridization of N-Grams okapi statistical measurements and a kernel function. In order to have a robust descriptor index, he used a semantic graph to model the semantic connections between terms with an auxiliary dictionary to increase the connectivity of the graph. First, the document is modeled by a graph. Then the graph is fed by a dictionary of concepts. The word weights are then calculated using a radial basis function. This has improved the performance of the indexing system. Zaki *et al.* adopt the k nearest neighbors as a method of classification and recall and precision as metrics of evaluation.

Liu *et al.* [13] propose a keyword extraction method based on semantic grouping which guarantees good semantic coverage of the document. The method makes it possible to extract the candidate terms which will be grouped in classes after having calculated the semantic links between these terms. This grouping consists in developing a set of reference words for each class. The reference words are used to extract the keywords after filtering the candidate terms.

3. Proposed System of Contextualization

Fig. 1. System overview.

Most of the researches which are done in extracting unsupervised information focuses on the extraction of keywords, and few works offer methods for extracting the semantic relation between them. In this paper, we present a new approach for indexing text documents based on context discovery. Clustering methods are the main idea that leads to this approach. It is about grouping together sentences issued from the same document, and expressing a semantic proximity with each other. Our approach is composed of three steps:

- **Segmentation:** The first step is to define the text's unit that will be taken into consideration to form contexts.

- Clustering: The second step is gathering semantically close sentences onto clusters. In our experiment, we use iterative K-means with Euclidian distance as a metric.
- Building Contexts: The third step is grouping the obtained sentences into clusters. Hence, each cluster represent a context. Each document will obviously have at least one context.

3.1. Segmentation: Sentence Splitting

Sentence splitting is the process of dividing textual documents into meaningful units, that we consider as context. This is an intermediate process. Humans understand the sentence when reading text, and computers implements artificial methods to split meaningful sentences, which are the subject of our natural language processing. Texts have explicit sentence boundary markers which is punctuation.

Hence, we use punctuation signs to delimit a sencece. Since Arabic language do not have capital letters, our approach of splitting sentences is based on full stop, exclamation and interrogation points (".", "!", "?") cutting.

3.2. Context Building

In order to enable an effective clustering process, the word frequencies need to be normalized in terms of their relative frequency of presence in the document and over the entire collection. In general, a common representation used for text processing is the *vector space based* TF-IDF representation. In the TF-IDF representation, the term frequency for each word is normalized by the *inverse document frequency*, or IDF. The inverse document frequency normalization reduces the weight of terms which occur more frequently in the collection. This reduces the importance of common terms in the collection, ensuring that the matching of documents be more influenced by that of more discriminative words which have relatively low frequencies in the collection (Fig. 1).

In addition, a sub-linear transformation function is often applied to the term frequencies in order to avoid the undesirable dominating effect of any single term that might be very frequent in a document. Text clustering algorithms are divided into a wide variety of different types such as agglomerative clustering algorithms, partitioning algorithms etc.

Different clustering algorithms have different tradeoffs in terms of effectiveness and efficiency.

In this work, we used iterative K-means as method of clustering. The advantage of this method is that it allocate the optimal K to each document.

3.3. Document Modeling

Fig. 2.Space Vector Modeling for a simple and a contextualized document.

To model the document we use the Space Vector Model. After forming contexts, we recalculate the new

score of each word according to the context. Thus, each document can be represented by a set of vectors, each vector modelise a given context (Fig. 2).

Then, for classification aim, every document is modeled by one vector wich is the must powerful vector (the vector with big scores). In other words, the most powerful vector represent the dominant context in a text.

<p>وتندرج المحطة، في إطار المخطط الأزرق، الذي يشكل الحجر الأساسي لما يعرف بـ "روية 2010" السياسية، الأمطار تلهب أسعار الخضر والفواكه، يقبل أقل من أسبوع بدأت أسعار الكثير من الخضر والفواكه، خصوصا منها الأكثر استهلاكاً، في الارتفاع بفعل نزول الأمطار في العديد من المناطق المغربية، وهي زيادة تلقائية ناتجة عن الصعوبة الكبيرة التي يواجهها العامل الزراعيون في قطف واستخراج الخضر والفواكه من الضيعات، وهو ما ينتج عن تناقص الكميات المعروضة في أسواق الجملة والأسواق الموازية لها فترتفع الأسعار. ويوضح محمد أفخشي رئيس الفدرالية الوطنية لوكلاء مستخدمي أسواق الجملة للخضر والفواكه في المغرب أن سعر الطماطم في بعض المناطق كفسل زاد ليصل إلى 6 دراهم بعدما كان أقل من 5 دراهم قبل أسابيع قليلة، في حين لا يتجاوز سعرها في سوق الجملة 3 دراهم، وأما البصل فيتراوح سعر الكيلوغرام الواحد منها في الأسواق الشعبية بين 4 و5 دراهم، وفي سوق الجملة بين 1.80 درهم ودرهمين، فيما يصل ثمن البطاطس إلى 5 دراهم لدى المستهلك ويتراوح بين درهمين و2.70 درهم في سوق الجملة. وواصل ثمن التفاح ارتفاعه بحيث يتراوح بين 12 و20 درهم للكيلوغرام في بعض الأسواق الشعبية، وهو ما يمثل ضعف الثمن تقريبا الذي يباع به في أسواق الجملة، فيما يناهز ثمن الكلمنتين 6 دراهم بزيادة درهمين تقريبا على أسواق الجملة، وهو من الفواكه التي يكثر الإقبال عليها في فترات الخريف والشتاء لوفرتها وفوائدها الصحية، في حين يصل ثمن الموز، الذي لا يتميز عن باقي الفواكه بوفرة على مدار أشهر السنة بثمن غير باهظ، إلى 11 و12 درهم للكيلوغرام. الزيادة في أسعار الخضر والفواكه لا ترجع فقط إلى عامل صعوبة إخراج هذه المنتجات من الأراضي الفلاحية، إن تأثير البرودة على الدورة الحياتية لنمو بعض الخضر والفواكه يظهر جليا من خلال تباطي نضجها مما ينقص الكمية المتوفرة في الأسواق الشعبية بين 4 و5 دراهم، وفي سوق الجملة بين 1.80 درهم ودرهمين، فيما يصل ثمن البطاطس إلى 5 دراهم لدى المستهلك ويتراوح بين درهمين و2.70 درهم في سوق الجملة. وواصل ثمن التفاح ارتفاعه بحيث يتراوح بين 12 و20 درهم للكيلوغرام في بعض الأسواق الشعبية، وهو ما يمثل ضعف الثمن تقريبا الذي يباع به في أسواق الجملة، فيما يناهز ثمن الكلمنتين 6 دراهم بزيادة درهمين تقريبا على أسواق الجملة، وهو من الفواكه التي يكثر الإقبال عليها في فترات الخريف والشتاء لوفرتها وفوائدها الصحية، في حين يصل ثمن الموز، الذي لا يتميز عن باقي الفواكه بوفرة على مدار أشهر السنة بثمن غير باهظ، إلى 11 و12 درهم للكيلوغرام. إن تأثير البرودة على الدورة الحياتية لنمو بعض الخضر والفواكه يظهر جليا من خلال تباطي نضجها مما ينقص الكمية المتوفرة في الأسواق الشعبية بين 4 و5 دراهم، وفي سوق الجملة بين 1.80 درهم ودرهمين، فيما يصل ثمن البطاطس إلى 5 دراهم لدى المستهلك ويتراوح بين درهمين و2.70 درهم في سوق الجملة. وواصل آخر لا يقل أهمية يسهم في رفع أثمان الخضر والفواكه عند وصولها إلى أسواق الجملة ثم إلى الأسواق الشعبية، وهو غلاء سعر نقل هذه البضائع، ويستل الممتدح السابق على نموذج شاحنة تنقل كميات من الكلمنتين من مدينة بركان، المشهورة بإنتاج الليمون، ومدينة فاس، بحيث ترتفع كلفة نقل حمولة الشاحنة من هامش يتراوح بين 1700 و2000 درهم إلى 3000 و3500 درهم.</p>	<p>ترجع القطاعات المصدرة تؤكد التوقعات أن الوضعية الاقتصادية السائدة على الصعيد الدولي، سستمر في التأثير، حتما، على أداء الاقتصاد الوطني، خاصة على مستوى القطاعات المصدرة للخدمات والسلع، وقطاع السياحة، وتحويلات الجالية المغربية المقيمة بالخارج، فضلا عن تراجع الاستثمارات الخارجية، خصوصا على مستوى السياحة، والعقار. وكثت الإحصائيات الأخيرة لمكتب الصرف، أفادت أن الظرفية ذاتها، انعكست سلبا على أداء الميزان التجاري، إذ أن العجز بلغ، خلال الأشهر التسعة الأولى من 2009، ما مجموعه 109.8 ملايين درهم، مسجلا تراجعا بـ 12.8 في المائة، مقارنة بالمستوى المسجل، سنة من قبل. وحسب المصدر، بلغت صادرات البضائع في مجملها، عند نهاية شتير 2009 حوالي 83.6 مليار درهم، مقابل 126.6 مليار درهم، سنة من قبل، مسجلة بدورها تراجعا بنسبة 34 في المائة، في حين ارتفعت الواردات إلى 193.5 درهما، أي بانخفاض بلغ 23.4 في المائة. وعزا مكتب الصرف أسباب تراجع الصادرات، أساسا، إلى تراجع مبيعات الفوسفات ومشتقاته، والمواد الإلكترونية، والألبسة، والجوارب، والمواد الغذائية، من بينها الحوامض، والخضر، والفواكه، والقمريات، والرخويات، والصفديات. وأكد المصدر، أنه جرى تسجيل التوجه نفسه نحو التراجع، بالنسبة إلى المنتجات المستوردة، خاصة ما يتعلق بالمنتجات الطاقية، ومنها النفط الخام، والمواد الغذائية، بسبب تأثير تراجع مشتريات الحبوب، والمواد الاستهلاكية، ومواد التجهيز. من جهتها، سستمر تنقذات الاستثمارات الخارجية المباشرة إلى المغرب بالانخفاض، للسنة الثانية على التوالي، وسجلت الاستثمارات انخفاضاً بنسبة 20 في المائة، سنة 2009، بعدما سجلت تراجعا بنسبة 37 في المائة سنة 2008، إثر تسجيل رقم قياسي، من حيث القيمة، بلغت حوالي 37 مليار درهم، سنة 2007. ومن أبرز الاستثمارات، التي فقدها المغرب بسبب الأزمة، الاستثمار، إذ كان من المقرر أن تنجزه المجموعة الأميركية - "كولوني كابيتال"، لإقامة وحدات فندقية راقية، في محطة تغازوت، على بعد 15 كيلومترا من مدينة أكادير. وتندرج المحطة، في إطار المخطط الأزرق، الذي يشكل الأمطار تلهب أسعار الخضر والفواكه الحجر الأساسي لما يعرف بـ "روية 2010" السياسية.</p>
<p>ترجع القطاعات المصدرة تؤكد التوقعات أن الوضعية الاقتصادية السائدة على الصعيد الدولي، سستمر في التأثير، حتما، على أداء الاقتصاد الوطني، خاصة على مستوى القطاعات المصدرة للخدمات والسلع، وقطاع السياحة، وتحويلات الجالية المغربية المقيمة بالخارج، فضلا عن تراجع الاستثمارات الخارجية، خصوصا على مستوى السياحة، والعقار. وكثت الإحصائيات الأخيرة لمكتب الصرف، أفادت أن الظرفية ذاتها، انعكست سلبا على أداء الميزان التجاري، إذ أن العجز بلغ، خلال الأشهر التسعة الأولى من 2009، ما مجموعه 109.8 ملايين درهم، مسجلا تراجعا بـ 12.8 في المائة، مقارنة بالمستوى المسجل، سنة من قبل. وحسب المصدر، بلغت صادرات البضائع في مجملها، عند نهاية شتير 2009 حوالي 83.6 مليار درهم، مقابل 126.6 مليار درهم، سنة من قبل، مسجلة بدورها تراجعا بنسبة 34 في المائة، في حين ارتفعت الواردات إلى 193.5 درهما، أي بانخفاض بلغ 23.4 في المائة. وعزا مكتب الصرف أسباب تراجع الصادرات، أساسا، إلى تراجع مبيعات الفوسفات ومشتقاته، والمواد الإلكترونية، والألبسة، والجوارب، والمواد الغذائية، من بينها الحوامض، والخضر، والفواكه، والقمريات، والرخويات، والصفديات. وأكد المصدر، أنه جرى تسجيل التوجه نفسه نحو التراجع، بالنسبة إلى المنتجات المستوردة، خاصة ما يتعلق بالمنتجات الطاقية، ومنها النفط الخام، والمواد الغذائية، بسبب تأثير تراجع مشتريات الحبوب، والمواد الاستهلاكية، ومواد التجهيز. من جهتها، سستمر تنقذات الاستثمارات الخارجية المباشرة إلى المغرب بالانخفاض، للسنة الثانية على التوالي، وسجلت الاستثمارات انخفاضاً بنسبة 20 في المائة، سنة 2009، بعدما سجلت تراجعا بنسبة 37 في المائة سنة 2008، إثر تسجيل رقم قياسي، من حيث القيمة، بلغت حوالي 37 مليار درهم، سنة 2007. ومن أبرز الاستثمارات، التي فقدها المغرب بسبب الأزمة، الاستثمار، إذ كان من المقرر أن تنجزه المجموعة الأميركية - "كولوني كابيتال"، لإقامة وحدات فندقية راقية، في محطة تغازوت، على بعد 15 كيلومترا من مدينة أكادير.</p>	<p>عامل آخر لا يقل أهمية يسهم في رفع أثمان الخضر والفواكه عند وصولها إلى أسواق الجملة ثم إلى الأسواق الشعبية، وهو غلاء سعر نقل هذه البضائع، ويستل الممتدح السابق على نموذج شاحنة تنقل كميات من الكلمنتين من مدينة بركان، المشهورة بإنتاج الليمون، ومدينة فاس، بحيث ترتفع كلفة نقل حمولة الشاحنة من هامش يتراوح بين 1700 و2000 درهم إلى 3000 و3500 درهم.</p>

Fig. 3. Original text.

<p>وتندرج المحطة، في إطار المخطط الأزرق، الذي يشكل الحجر الأساسي لما يعرف بـ "روية 2010" السياسية، الأمطار تلهب أسعار الخضر والفواكه، يقبل أقل من أسبوع بدأت أسعار الكثير من الخضر والفواكه، خصوصا منها الأكثر استهلاكاً، في الارتفاع بفعل نزول الأمطار في العديد من المناطق المغربية، وهي زيادة تلقائية ناتجة عن الصعوبة الكبيرة التي يواجهها العامل الزراعيون في قطف واستخراج الخضر والفواكه من الضيعات، وهو ما ينتج عن تناقص الكميات المعروضة في أسواق الجملة والأسواق الموازية لها فترتفع الأسعار. ويوضح محمد أفخشي رئيس الفدرالية الوطنية لوكلاء مستخدمي أسواق الجملة للخضر والفواكه في المغرب أن سعر الطماطم في بعض المناطق كفسل زاد ليصل إلى 6 دراهم بعدما كان أقل من 5 دراهم قبل أسابيع قليلة، في حين لا يتجاوز سعرها في سوق الجملة 3 دراهم، وأما البصل فيتراوح سعر الكيلوغرام الواحد منها في الأسواق الشعبية بين 4 و5 دراهم، وفي سوق الجملة بين 1.80 درهم ودرهمين، فيما يصل ثمن البطاطس إلى 5 دراهم لدى المستهلك ويتراوح بين درهمين و2.70 درهم في سوق الجملة. وواصل ثمن التفاح ارتفاعه بحيث يتراوح بين 12 و20 درهم للكيلوغرام في بعض الأسواق الشعبية، وهو ما يمثل ضعف الثمن تقريبا الذي يباع به في أسواق الجملة، فيما يناهز ثمن الكلمنتين 6 دراهم بزيادة درهمين تقريبا على أسواق الجملة، وهو من الفواكه التي يكثر الإقبال عليها في فترات الخريف والشتاء لوفرتها وفوائدها الصحية، في حين يصل ثمن الموز، الذي لا يتميز عن باقي الفواكه بوفرة على مدار أشهر السنة بثمن غير باهظ، إلى 11 و12 درهم للكيلوغرام. الزيادة في أسعار الخضر والفواكه لا ترجع فقط إلى عامل صعوبة إخراج هذه المنتجات من الأراضي الفلاحية، إن تأثير البرودة على الدورة الحياتية لنمو بعض الخضر والفواكه يظهر جليا من خلال تباطي نضجها مما ينقص الكمية المتوفرة في الأسواق الشعبية بين 4 و5 دراهم، وفي سوق الجملة بين 1.80 درهم ودرهمين، فيما يصل ثمن البطاطس إلى 5 دراهم لدى المستهلك ويتراوح بين درهمين و2.70 درهم في سوق الجملة. وواصل ثمن التفاح ارتفاعه بحيث يتراوح بين 12 و20 درهم للكيلوغرام في بعض الأسواق الشعبية، وهو ما يمثل ضعف الثمن تقريبا الذي يباع به في أسواق الجملة، فيما يناهز ثمن الكلمنتين 6 دراهم بزيادة درهمين تقريبا على أسواق الجملة، وهو من الفواكه التي يكثر الإقبال عليها في فترات الخريف والشتاء لوفرتها وفوائدها الصحية، في حين يصل ثمن الموز، الذي لا يتميز عن باقي الفواكه بوفرة على مدار أشهر السنة بثمن غير باهظ، إلى 11 و12 درهم للكيلوغرام. إن تأثير البرودة على الدورة الحياتية لنمو بعض الخضر والفواكه يظهر جليا من خلال تباطي نضجها مما ينقص الكمية المتوفرة في الأسواق الشعبية بين 4 و5 دراهم، وفي سوق الجملة بين 1.80 درهم ودرهمين، فيما يصل ثمن البطاطس إلى 5 دراهم لدى المستهلك ويتراوح بين درهمين و2.70 درهم في سوق الجملة. وواصل آخر لا يقل أهمية يسهم في رفع أثمان الخضر والفواكه عند وصولها إلى أسواق الجملة ثم إلى الأسواق الشعبية، وهو غلاء سعر نقل هذه البضائع، ويستل الممتدح السابق على نموذج شاحنة تنقل كميات من الكلمنتين من مدينة بركان، المشهورة بإنتاج الليمون، ومدينة فاس، بحيث ترتفع كلفة نقل حمولة الشاحنة من هامش يتراوح بين 1700 و2000 درهم إلى 3000 و3500 درهم.</p>	<p>ترجع القطاعات المصدرة تؤكد التوقعات أن الوضعية الاقتصادية السائدة على الصعيد الدولي، سستمر في التأثير، حتما، على أداء الاقتصاد الوطني، خاصة على مستوى القطاعات المصدرة للخدمات والسلع، وقطاع السياحة، وتحويلات الجالية المغربية المقيمة بالخارج، فضلا عن تراجع الاستثمارات الخارجية، خصوصا على مستوى السياحة، والعقار. وكثت الإحصائيات الأخيرة لمكتب الصرف، أفادت أن الظرفية ذاتها، انعكست سلبا على أداء الميزان التجاري، إذ أن العجز بلغ، خلال الأشهر التسعة الأولى من 2009، ما مجموعه 109.8 ملايين درهم، مسجلا تراجعا بـ 12.8 في المائة، مقارنة بالمستوى المسجل، سنة من قبل. وحسب المصدر، بلغت صادرات البضائع في مجملها، عند نهاية شتير 2009 حوالي 83.6 مليار درهم، مقابل 126.6 مليار درهم، سنة من قبل، مسجلة بدورها تراجعا بنسبة 34 في المائة، في حين ارتفعت الواردات إلى 193.5 درهما، أي بانخفاض بلغ 23.4 في المائة. وعزا مكتب الصرف أسباب تراجع الصادرات، أساسا، إلى تراجع مبيعات الفوسفات ومشتقاته، والمواد الإلكترونية، والألبسة، والجوارب، والمواد الغذائية، من بينها الحوامض، والخضر، والفواكه، والقمريات، والرخويات، والصفديات. وأكد المصدر، أنه جرى تسجيل التوجه نفسه نحو التراجع، بالنسبة إلى المنتجات المستوردة، خاصة ما يتعلق بالمنتجات الطاقية، ومنها النفط الخام، والمواد الغذائية، بسبب تأثير تراجع مشتريات الحبوب، والمواد الاستهلاكية، ومواد التجهيز. من جهتها، سستمر تنقذات الاستثمارات الخارجية المباشرة إلى المغرب بالانخفاض، للسنة الثانية على التوالي، وسجلت الاستثمارات انخفاضاً بنسبة 20 في المائة، سنة 2009، بعدما سجلت تراجعا بنسبة 37 في المائة سنة 2008، إثر تسجيل رقم قياسي، من حيث القيمة، بلغت حوالي 37 مليار درهم، سنة 2007. ومن أبرز الاستثمارات، التي فقدها المغرب بسبب الأزمة، الاستثمار، إذ كان من المقرر أن تنجزه المجموعة الأميركية - "كولوني كابيتال"، لإقامة وحدات فندقية راقية، في محطة تغازوت، على بعد 15 كيلومترا من مدينة أكادير.</p>
<p>عامل آخر لا يقل أهمية يسهم في رفع أثمان الخضر والفواكه عند وصولها إلى أسواق الجملة ثم إلى الأسواق الشعبية، وهو غلاء سعر نقل هذه البضائع، ويستل الممتدح السابق على نموذج شاحنة تنقل كميات من الكلمنتين من مدينة بركان، المشهورة بإنتاج الليمون، ومدينة فاس، بحيث ترتفع كلفة نقل حمولة الشاحنة من هامش يتراوح بين 1700 و2000 درهم إلى 3000 و3500 درهم.</p>	<p>عامل آخر لا يقل أهمية يسهم في رفع أثمان الخضر والفواكه عند وصولها إلى أسواق الجملة ثم إلى الأسواق الشعبية، وهو غلاء سعر نقل هذه البضائع، ويستل الممتدح السابق على نموذج شاحنة تنقل كميات من الكلمنتين من مدينة بركان، المشهورة بإنتاج الليمون، ومدينة فاس، بحيث ترتفع كلفة نقل حمولة الشاحنة من هامش يتراوح بين 1700 و2000 درهم إلى 3000 و3500 درهم.</p>

Fig. 4. Text after contextualization, attributed to class Economics.

4. Experiments and Results

To evaluate our approach, we have used K Nearest Neighbor as a classifier. First, we conduct a test with the standard TFIDF system that represent a word by its weight. Then, we have attempted to classify each document by its most powerful vector extracted from the most dominant context.

4.1. Corpus

To validate this new approach, we tested it on a varied corpus of 124 documents representing scripts of various lengths. Each article belong either to Economics, Politics or Sport.

Our experiments are at their beginning. So far, we have tested our system on a small database to approve its effectiveness. However, we develop a more complete corpus to test the efficiency of our coming algorithms.

4.2. Classification Results

Text classification is an important part of the text mining process. It is providing a set of training data

(tagged documents) to the classification system. The task then is to determine a classification model that is able to affect the right class for a new document.

To evaluate the classification performance, three metrics are used: precision, recall and F-measure.

Table 1. Classification Results

	Precision	Recall	F-measure
TF IDF indexation	0.619	0.6	0.609
Contextual indexation	0.805	0.8	0.802

Table 1 shows different results for each measure. These results are expressed through the recall, precision and F-measure criteria. In particular, they show the relevance of using contextual indexation which greatly improves the measures' performances.

The first method (TF IDF) is basic, it represents each document by a vector of characteristics coming from the frequency weighting. However, in our proposed approach, after carrying out all the contextualization steps, the system provides a number of vectors that correspond to the number of contexts per document. Then, our system represents each document by the most significant vector that corresponds to the most dominant context in a given document.

4.3. Discussion

Compared to other languages, the Arabic language has a very rich morphological variation and extremely complex syntactic and flexional characteristics, which is one of the main reasons why the lack of research methods in the field of treatment of Arabic. Indexing and classification of texts are important tasks of text mining process. A typical process of text classification consists of the following steps: pre-processing, indexing, dimension reduction, and classification.

Statistical approaches represent the text a list of weighted keywords. This representation is adapted to capture the frequency of occurrence of a word and ignores the structural and semantic information of the document. However, it is not suitable for modeling semantics. This type of technique offers good results in specific cases, However, taking into account the semantics, it is clear that the approach we have proposed improves the classification results.

Experimental evaluation of the classifier is the final step of the indexing process. It usually tries to evaluate the effectiveness of a classifier, namely its ability to make categorization decisions. There are many measures for this purpose, each highlighting a particular property of the system. The documents processing in the Arabic language is confronted with another problem, that of the evaluations of the methods on the corpus. In most works on Arabic texts, and in the absence of a free standard corpus, authors construct their own corpus. They choose the number of categories and their themes. For each category the documents are collected manually. Documents belonging to several categories are eliminated. However, in order to test the accuracy of the different methods, they must be applied to the same corpus, even more so that a method proves its effectiveness it must be applied to several corpus of different themes.

5. Conclusion

The semantic proximity between words must be highlighted when we deal with complex and unstructured documents such as texts in Arabic. For this purpose, it is essential to broaden our reflection to the adapted representation models to the nature of our resources. For this aim, we studied the research model based on the contextual indexation model.

The integration of a semantic measure between sentences in this approach is needed. For this reason, we

have introduced our clustering contribution to formalize the adaptation of the model based on the semantic proximity. The advantage of this model is that it does not need a preliminary glossary to identify terms in order to assign them a weight, since the identification of terms is made simply from a document processing.

References

- [1] EL Bazzi, M. S., & Mammass, D. (2017). A graph based method for Arabic document indexing. *Proceedings of 2017 International Conference on Information and Digital Technologies (IDT)* (pp. 308-312).
- [2] Mesleh, A. (2007). Support vector machines based Arabic language text classification system: Feature selection comparative study. *Proceedings of the 12th WSEAS International Conference on Applied Mathematics* (pp. 11-16).
- [3] Thabtah F., Eljinini, M., Zamzeer, M., & Hadi, W. (2009). Naïve bayesian based on Chi square to categorize Arabic data. *Proceedings of the 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, IBIMA'2009* (pp. 930-935).
- [4] Bawaneh, M. J., Alkoffash, M. S., & Al Rabea, A. I. (2008). Arabic text classification using K-NN and Naive bayes. *Journal of Computer Science*, 4, 600-605.
- [5] Kanaan G., Yaseen, M., Al-Shalabi, R., Al-Sarayreh, B., & Mustafa, A. (2009). Using EM for text classification on Arabic. *Proceedings of the Second International Conference on Arabic Language Resources and Tools*.
- [6] Raheel, S., & Dichy, J. (2010). An empirical study on the feature qs type effect on the automatic classification of Arabic documents. *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'10* (pp. 673-686).
- [7] Zaki, T., Mammass, D., & Ennaji, A. (2010). A semantic proximity based system of Arabic text indexation. *International Conference on Image and Signal Processing (ICISP)*.
- [8] Al-Salemi, B., & Aziz, M. J. A. (2011). Statistical Bayesian learning for automatic Arabic text categorization. *Journal of Computer Science*, 7(1), 39-45.
- [9] Mansour N., Haraty, R. A., Daher, W., & Houri, M. (2008). An auto-indexing method for Arabic text. *Information Processing and Management*, 44(4), 1538-154.
- [10] Tomokiyo, T., & Hurst, M. (2003). A language model approach to keyphrase extraction. *Proceedings of the ACL Workshop on Multiword Ex- Pressions*.
- [11] Jamoussi, S. (2009). Une nouvelle représentation vectorielle pour la classification sémantique. *TAL*, 50.
- [12] Zaki, T., Mammass, D., Ennaji, A., & Nicolas, S. (2014). A kernel hybridization NGram-Okapi for indexing and classification of Arabic documents. *Journal of Information and Computing Science*, 9(2), 141-153.
- [13] Liu, Z., Peng, L., Yabin, Z., & Maosong, S. (2009). Clustering to find exemplar terms for keyphrase extraction. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 257-266).

Mohamed Salim EL BAZZI received the master's degree in computer sciences from Ibn Zohr University, and currently is a Ph.D student at the same university. He is a member of " images pattern recognition — intelligent and communicating systems " Laboratory. He is working on text mining under the supervision of Prof. Driss Mammass. His research areas are systems of information, text indexing, classification, data mining, opinion mining, natural language processing and information retrieval.

Driss Mammass is professor of higher education at the Faculty of Sciences, University Ibn Zohr, Agadir Morocco. He received a doctorat in mathematics in 1988 from Paul Sabatier University (Toulouse - France) and a doctorat d'Etat-es-sciences degrees in mathematics and image processing from University Ibn Zohr Agadir Morocco, in 1999. He supervises several Ph.D theses in the various research themes of mathematics and computer science. He used to be director of High School of Technology Agadir and the head of IRF-SIC laboratory.

Abdelati Ennaji has been an associate professor at the University of Rouen since 1993. He received his Ph.D from the University of Rouen in 1993 in the fields of machine learning and pattern recognition. His major scientific interest include incremental technics for statistical and hybrid machine learning, data analysis and clustering. The main applications of these activities concern pattern recognition problems and Arabic text mining and recognition.

Taher Zaki received the Ph.D degree in computer science from Ibn Zohr University, and now is a professor at the same University. He is a member of "images pattern recognition — intelligent and communicating systems" laboratory. His research interests systems of information retrieval, text indexing and archive of document.