# *Enhancing Retrieval and Novelty Detection for Arabic Text Using Sentence level Information Pattern*


Esra'a AL-Shdaifat
Software Engineering Dept.
IT & CS Faculty
The Hashemite University
13115 Zarqa – Jordan
esraa_shdaifat@hu.edu.jo

Mohammed N. Al-Kabi
CIS Department
IT & CS Faculty
Yarmouk University
21163 Irbid – Jordan
mohammedk@yu.edu.jo

Emad Al-Shawakfa
CIS Department
IT & CS Faculty
Yarmouk University
21163 Irbid – Jordan
shawakfa@yu.edu.jo

Abdulla H Wahbeh
College of Business and
Information System
Dakota State University
Madison, SD, USA
ahwahbeh@pluto.dsu.edu



***Abstract*—Novelty detection is already used in many Natural Processing Language (NLP) applications, such as information retrieval systems, Web search engines, text summarization, question answering systems …etc. This study aims to detect novel Arabic sentence level information patterns. The Length Adjusted (LA) model is based on sentence level information patterns is used, which depends on the sentence length. Test results show a significant improvement in the performance of novelty detection for Arabic texts in terms of precision at top ranks.**



***Keywords; Novelty detection; information patterns; Information retrieval.***


## I. INTRODUCTION

Information Retrieval (IR) is an area or an art of study dedicated to search for information within textual documents, where the information needs expressed as queries submitted to an information retrieval system in order to provide its users with relevant information. It is also possible to submit an unstructured query or unstructured topic statement to the system. Therefore, there is a need at the present time for effective methods to produce intelligent information retrieval systems that are capable to deal with unstructured queries and documents effectively. The world has witnessed these days an explosion in the amount of unstructured data, especially, over the Internet [1]. Information retrieval systems (IRS) and digital libraries are based on the assumption that users have some knowledge about the information they are searching for. Also, there are a significant number of evidences that prove that users obtain information that they have no idea about at all [2].

With the continuing growth of information, users of information retrieval system need to get useful information quickly and without examining a lot of redundant information. Using the novelty detection will reduce the amount of redundant and non-relevant materials presented to the user.

The novelty detection has a significant impact on information retrieval systems, where the selective dissemination of information (SDI); used mainly in libraries and information sciences and denotes tools and resources used to inform searchers mainly about new information, and enable them to be always updated when they search for any topic. Shared interests of users' community and similarity to a topical profile are used by information filtering systems and most of SDI [3]. Recent efforts have looked at the retrieval of specific novel information. The goal of studies on novelty detection is to find techniques to reduce redundant and non-relevant material produced and presented by document retrieval systems. The ranked list of documents generated by document retrieval systems; which adapt novelty detection, usually filter out all non-relevant information, after which the residual essential relevant information will be scanned in search for old material to be discarded too. Afterward, document retrieval systems present the user with a ranked list of non-redundant relevant sentences. Usually, researchers start with a known set of relevant documents to simplify the novelty detection, and then facilitate the search for novel documents. The assumption is presumably that the process of finding relevant materials can be explored separately [4].

Up to the knowledge of the researchers; no such study was performed on Arabic text before and this would constitute the first of a kind in this area.

This paper is organized as follows: the second section will introduce related works, the third section presents the adopted methodology, the fourth section presents the results and evaluation, and the fifth section presents the conclusion and future work.

## II. RELATED WORK

Zhang, Callan, and Minka [5] extended an adaptive information filtering system to make decisions about the novelty and redundancy of relevant documents. Also, they have proposed a set of five redundancy measures with and without redundancy thresholds. The results of the conducted experiments by this team have proven that the cosine similarity measure and a redundancy measure based on a mixture of language models are effective techniques to identify the redundant documents.

Allan, Wade, and Bolivar [4] have presented a Topic Detection and Tracking (TDT) research and evaluation project; which is dedicated to novel online event detection and tracking. TDT tasks interested in inter-topic or inter-event novelty detection, in order to determine whether two news stories cover the same occasion. TDT is interested in with story-level online evaluation, where the news stories are presented one after another to be evaluated sequentially, and identify the new news stories. This task was presented more satisfactorily in a research by [6] on temporal summarization; where the main concern of their effort was to develop a useful evaluation model.

Yang, Pierce, and Carbonell [7] have also proposed the usage of clustering techniques to detect different events. The task of the proposed system is to detect automatically new events from a temporally ordered stream of news stories, either retrospectively or as the stories arrive. By applying hierarchical and non-hierarchical document clustering algorithms, they found that temporal distribution patterns of document clusters provided useful information for improvement in both retrospective detection and on-line detection of novel events.

Allan, Papka, and Lavrenko [8] have also described new event detection and event tracking system within a stream of broadcast news stories. The system has to make decisions about a single story before looking at subsequent stories. This approach uses a single pass clustering algorithm with a novel threshold model that includes the properties of events as a main component.

Discovering new events automatically from chronologically ordered documents is a real challenge to researchers in this field. A particular form of novelty detection called First Story Detection (FSD); which is known as the most difficult in the field of Topic Detection and Tracking (TDT). FSD aims to detect on-line new news stories as soon as they arrive in the sequence of documents.

Yang, Pierce, and Carbonell [9] have proposed a new supervised learning algorithm to classify on-line documents by topic, and topic-conditioned feature weights to measure the novelty of documents within a topic at the event level. In addition, the authors have focused on using named-entities for event-level novelty detection and using feature-based heuristics extracted from the topic histories. The results of the tests show a substantial improvement over the traditional one-level approach to detect novel documents.

Fernández, and Losada [10] have addressed Local Context Analysis (LCA) to detect new and relevant sentences within documents related to a certain topic. LCA is beneficial to researchers in different areas of study; such as text summarization, information retrieval, Web search engines, question answering, etc. The core idea of this method is based on a common term from the top-ranked relevant documents that tend to co-occur with query terms within the top-ranked documents.

A significant related work proposed by [11]. Starting from providing a new definition of novelty as "new answers to the potential questions representing a user's request or information need", they have suggested new novelty detection approach that was based on the identification of sentence level information patterns. In the first step, the user query is transformed into a potential question(s), in order to identify correspondent query-related information patterns that contain query terms and the required answer types [12] [13]. In the second step, the new information is extracted through detecting sentences that include unseen previously answers relevant to the query-related patterns. The information–pattern-based novelty detection (IPND) suggested by [11], depends on three sentence information patterns that include sentence length, opinion patterns and named entities. The sentence length represents the number of words in the sentence.

## III. METHODOLOGY

This section discusses the proposed Information-Pattern-based Novelty Detection (IPND) approach; which is illustrated in Figure 1. Two important steps demonstrated in this approach namely; relevant sentence retrieval and novel sentence extraction.

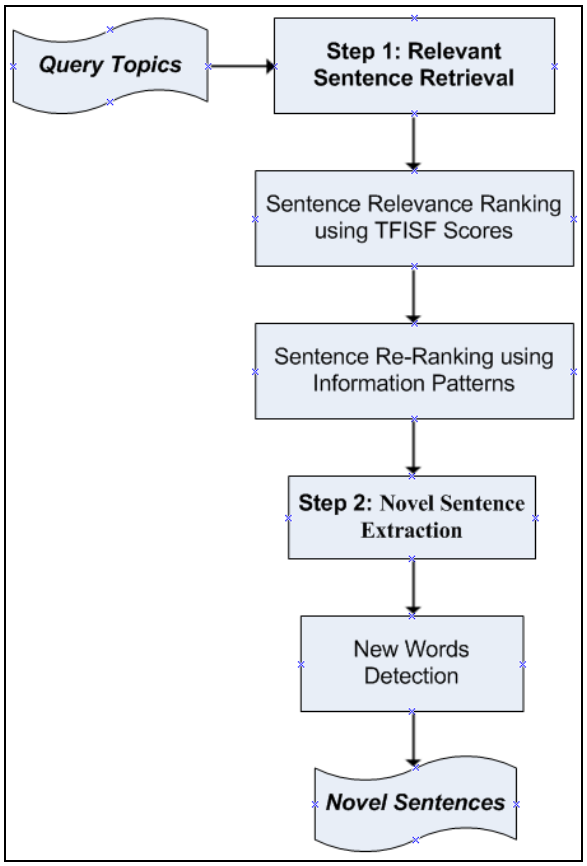


Figure 1. Information-Pattern-based Novelty Detection (IPND) approach.

### A. *Relevant Sentence Retrieval*

The task of the relevant sentence retrieval module is to retrieve sentences that are relevant to a user information need (Query). It first takes query words and searches in a data collection to retrieve sentences that are topically relevant to the query (initial ranking). It then re-ranks the retrieved sentences

using the sentence information patterns, including the sentence length only. Sentences that do not satisfy the query are filtered out because they are unlikely to have potential answers to a user information need. Fig. 2 illustrates the overall ranking process.

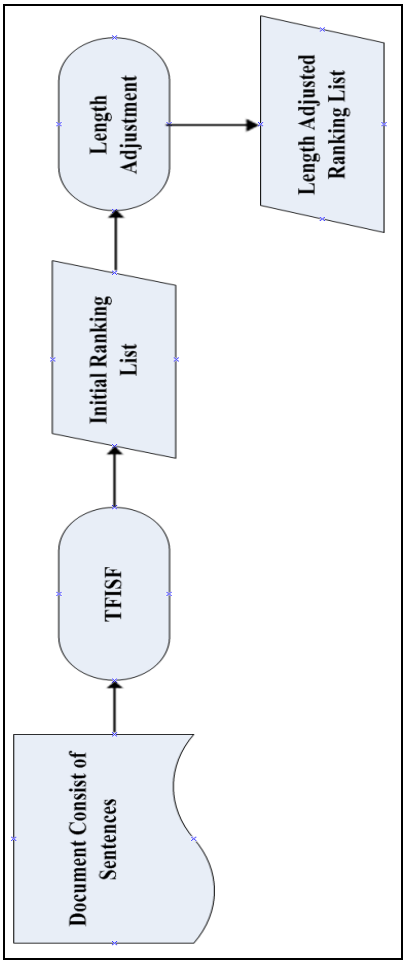


Figure 2. Ranking process.

*1) Initial Ranking Using TFISF Model*

In this research the Term Frequency-Inverse Sentence Frequency (TFISF) model was used to determine relevant sentences for the initial relevance ranking score. This model also used in other systems and reported to be able to achieve equivalently better performance scores compared to other techniques in sentence retrieval [1]. The initial TFISF relevance ranking score; So for a sentence, it is calculated according to the following formula:

$$S_0 = \sum_{i=1}^{n} [tfs(t_i) \times tfq(t_i) \times (isf(t_i)^2)] \quad (1)$$

Where $n$ is the total number of terms, $tfs(t_i)$ is the frequency of term $t_i$ in the sentence, and $tfq(t_i)$ is the frequency of term $t_i$ in the query, $isf(t_i)$ is inverse sentence frequency (instead of inverse document frequency in document retrieval) [1]. The inverse sentence frequency is calculated as:

$$isf(t_i) = \log \frac{N}{N_{ti}} \quad (2)$$

Where $N$ is the total number of sentences in the collection, $N_{ti}$ is the total number of sentences that include the term $t_i$.

*B. Novel Sentence Extraction*

This study includes two ranking lists; the initial ranking list which results from applying (TFISF) model, and the LA ranking list; as given in Figure 2.

## IV. Results And Evaluation

This section demonstrates and discusses the main experimental results. A corpus was collected from five websites; we included in the corpus all available data collected from the covered websites on the collection date. Stop words elimination was performed on the whole corpus as a preparation step, depending on the revised stop word list mentioned in section three.

To evaluate the performance of the information pattern based models, the Term Frequency-Inverse Sentence Frequency (TFISF) model has served as the baseline for comparing the performance of relevant sentence retrieval for novelty detection. The evaluation measure used for performance comparison is precision at rank N (N = 3, 5, 7, 9, and 11). Precision at rank N was calculated for each ranking list to compare the ranking of them, using the following formula:

$$Precision = \frac{\text{Total no. of novel relevant retrieved sentences}}{\text{Total number of retrieved sentences}} \quad (3)$$

Note that precision at top ranks is useful in real applications where users only want to go through a small number of sentences [1]. After finding the precision at rank $N$ for different queries, the average precision at rank $N$ for each of the two models was calculated. Table 1 shows the results.

Table 1. Average precision at rank $N$ for TFISF model.

| Rank # | Average Precision |
|---|---|
| 3 | 0.7 |
| 5 | 0.56 |
| 7 | 0.57 |
| 9 | 0.48 |
| 11 | 0.5 |

As shown in Table 3, the LA model has achieved the lowest precision at rank 3 with a value of 0.60, and then this value has increased to 0.76, 0.82, and 0.84 for ranks 5, 7, and 9 respectively. However, the value has dropped down again to 0.79 for the rank value 11. So the value of the average precision at rank *N* was increasing as we moved from *N*=3 to *N*=9 then has dropped down at rank 11.

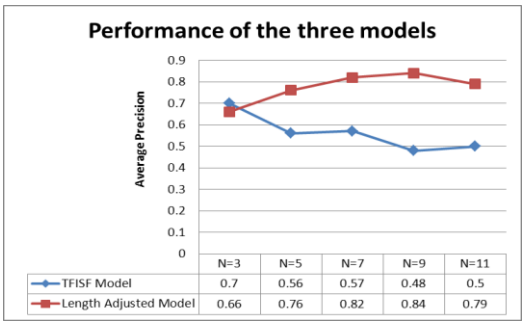


Figure 3. Average precision at rank *N* for the two models.

Fig. 3 shows a comparison of the results of the two models. We can notice that the average precision for the LA Model is much better than that of TFISF. On the other hand, the TFISF model has achieved a precision value at rank 3 better than the other model with a value of 0.70, however, the performance of the TFISF model has dropped down significantly for *N*=5 to *N*=11.

## V. Conclusion and Future Work

Novelty detection is an important approach used to identify new information, and reduce redundancy and the number of non-relevant information presented to users of systems; such as information retrieval systems, Web search engines, document filters, and cross-document summarization systems. Also, it can be used by different tasks of natural language processing (NLP); such as machine translation, text summarization, Question Answering, ...etc. This study aims to enhance the retrieval and novelty detection for Arabic text.

This research has proven that sentence level information patterns could be successfully applied to the Arabic text to improve the relevancy and novelty scores for Arabic retrieved sentences. The experiments have already proved that information patterns have a significant role in the novelty detection for general topics and relevance retrieval.

As a future research, we will test other models such as Like Opinion Adjusted (LOA) model to detect the novel information and compare it with the models used in this research.

## VI. References


[1] E. Greengrass, "Information Retrieval: A Survey," DOD Technical Report TR-R52-008-001, 2000.

[2] E. Toms, "Serendipitous Information Retrieval," In Proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland: European Research Consortium for Informatics and Mathematics, 2000.

[3] I. Soboroff, and D. Harman, "Novelty detection: the TREC experience," Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, p.105-112, October 06-08, 2005.

[4] J. Allan, C. Wade, and A. Bolivar, "Retrieval and novelty detection at the sentence level," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, July 28-August 01, 2003.

[5] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, August 11-15, 2002.

[6] J. Allan, R. Gupta, and V. Khandelwal, "Temporal summaries of new topics," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, p.10-18, September 2001.

[7] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, p.28-36, August 24-28, 1998.

[8] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, p.37-45, August 24-28, 1998.

[9] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, Topic-conditioned novelty detection, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, July 23-26, 2002.

[10] R. T. Fernández, and D. E. Losada, "Novelty detection using local context analysis," Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, July 23-27, 2007.

[11] X. Li, and W. B. Croft, "Improving novelty detection for general topics using sentence level information patterns," Proceedings of the 15th ACM international conference on Information and knowledge management, November 06-11, Arlington, Virginia, USA, 2006.

[12] X. Li, and W. B. Croft, "Evaluating question-answering techniques in Chinese," Proceedings of the first international conference on Human language technology research, p.1-6, San Diego, March 18-21, 2001.

[13] X. Li, "Syntactic features in question answering," Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, July 28-August 01, 2003.