Document Clustering for Large Arabic Language Corpora Using Unstructured Information Management Architecture with a Software Engineering Perspective

Basem Narmok
DePaul University
School of Computer Science
Amman, Jordan
narmok@ajniha.org

Abstract

The amount of data produced by humans is growing in a massively exponential way, especially after the Internet age; this is a real challenge for everyone seeking for knowledge in this huge amount of information with different languages! What this research will try to answer is how to retrieve knowledge from such “tsunami of data” [1] studying different challenges facing researchers dealing with unstructured information in a non-English language (in our case Arabic language), building on top of other researchers work and using the open source UIMA framework, toward Arabic document clustering, in the same time asking the question “how can we use this, or how to look at this in the eyes of a software engineer?” (e.g. Can we use such NLP technologies in requirements engineering?).

Key words: Arabic language, document clustering, natural language processing, software engineering, unstructured information, UIMA.

1. Introduction

"There is a tsunami of data that is crashing onto the beaches of the civilized world. This is a tidal wave of unrelated, growing data formed in bits and bytes, coming in an unorganized, uncontrolled, incoherent cacophony of foam. None of it is easily related, none of it comes with any organization methodology.

Now for the good news: There is a dune on the beach. There is a breakwater in the ocean that is clearly emerging in these last fleeting moments of the 20th century. The breakwater is indeed breaking up the tsunami of data and focusing it in a more organized way to answer our questions and concerns. There is a new breed of graphic designers, exhibition designers, illustrators and photographers, whose passion it is to make the complex clear." [1]

Human natural language is the main way of communication and knowledge transfer, and with the “tsunami of data” [1] we need new ways and tools to handle such a large content to reach the acquired knowledge, natural language processing isn't a magical set of tools to make the computer understand language, still it is a new field that needs more research and work to
mature, but it provides best ways/tools to extract concepts and to handle language, this research will study different NLP algorithms used in document clustering, for documents in Arabic language, using Apache open source framework for unstructured information management architecture (UIMA).

Now what is structured information? It is the “information whose intended meaning is unambiguous and explicitly represented in the structure or format of the data” [2], in the other hand unstructured information is the “information whose intended meaning is only loosely implied by the form” and “requires interpretation in order to approximate and extract its intended meaning” [2].

1.1 UIMA framework

The research will use the Unstructured Information Management Architecture (UIMA) which was developed by IBM and released as an open source in 2006, and it is now an Apache project, UIMA is a component framework for analyzing unstructured content such as text, audio, and video. Apache UIMA is Apache-licensed open source implementation that will be used to analyze large volumes of unstructured Arabic language information in order to discover knowledge; the focus will be on clustering such data in different categories.

![Figure 1. UIMA example application diagram from [2].](image)

Figure 1, shows an example of application built using UIMA framework, and it has two
distinguished phases: analysis and delivery. In the analysis the application can do document-level analysis or a collection-level analysis, the output of the analysis is a structured information, the structured information can be used in what called feedback-loop where it is used as a training set for the analysis of the other not-analyzed unstructured information, a good example of the feedback-loop is the “construction of ontologies where a collection of documents must first be analyzed to detect concepts”, in such case “looking at the entire collection improves the accuracy of concept determination” [2].

UIMA provides the capabilities to decompose the application into components, such components are required to be data-driven and self-descriptive in order to be reusable and interoperable (see figure1). UIMA was built with scalability in mind; it can scale to very large volumes by replicating processing pipelines over a cluster of networked nodes.

![UIMA application components (pipeline).](image)

UIMA provides a framework that can be used with unstructured text, audio, and video information, but in our research we will focus on unstructured text information.

### 1.2 Document clustering

Document clustering is very important when dealing with large amount of web content or documents as it will classify such large size of data in a set of groups or categories, this categorization is the big picture of the real content, having a big-picture makes handling large corpus of documents easier. “Document clustering is used in information retrieval and data mining to divide large unstructured corpora of documents into groups of more or less closely related documents”[3].

In this research document clustering will be built on top of UIMA, and will be used to cluster large corpora of Arabic content (usually such content is provided from the language data consortium (LDC), or collected from the web using special crawlers).

Processing a huge number of documents is useful in different ways, a good example for document clustering is news clustering, where we group news collected from different web sites or resources in a set of groups, this will ease the search in the news by topic, another example for processing a large set of documents and not clustering is to “discover evidence of probable
terrorist activities” [2].

1.3 Arabic language

Arabic is a Semitic language and ranked as the forth language by the number of native-language speakers (about 270 million speakers) so it comes after English language in usage, Arabic “language is written from right to left. It has very complex morphology, and the majority of words have a tri-letter root. The rest have either a quad-letter, penta-letter root or hexa-letter root.”[4].

The research will come over the characteristics of Arabic language and the differences between Arabic and English language in the light of natural language processing and document clustering.

1.4 Software engineering perspective

How can we use NLP technologies and UIMA framework in software engineering (e.g. Requirements Engineering)? After discussing how to cluster Arabic documents using UIMA framework, we will discuses how to apply this and use it in software engineering. What are the differences between Arabic language and English language in the light of software engineering, and what are the challenges in processing natural language for the use in software requirements engineering.

By knowing that “80 percent of all corporate information is unstructured” [2] and a large part of this information is related in a way or another to corporate different projects requirements, having the capability to classify and order that information is an important step toward requirements engineering.

2.0 Conclusions

The objectives of this research is to introduce UIMA framework architecture and to see if it is a useful framework for researchers in the fields of natural language processing and software engineering, the language of focus is the Arabic language as it is one of the largest languages used today in the same time the one with the least NLP research papers.

Document clustering in this research will use clustering algorithms from natural language processing and will discuss differences between the algorithms taking in consideration Arabic language requirements and challenges.

One of the major parts of this research is to discuss how can we use UIMA, and NLP in Software Engineering (e.g. Requirements Engineering), the research will introduce how the UIMA architecture can be used in software engineering (e.g. Maybe in building a requirements traceability engine).

It is a rich research that tries to bridge different research fields and technologies in a practical way toward handling and analyzing large sizes of data. Also it is rarely to find a mature research
related to Arabic language, and this research tries to contribute to the field of Arabic natural language processing, toward making the Arabic content (online/offline) accessible to the whole world of knowledge searchers.

References


