Journal of King Saud University – Computer and Information Sciences 29 (2017) 196–203

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Formal Concept Analysis for Arabic Web Search Results Clustering

Issam Sahmoudi *, Abdelmonaime Lachkar

Dept. of Electrical and Computer Engineering, ENSA, USMBA, Fez, Morocco

ARTICLE INFO

Article history:
Received 31 January 2016
Revised 30 June 2016
Accepted 19 September 2016
Available online 28 September 2016

Keywords:
Arabic language
Formal Concept Analysis
Web Search Results Clustering

ABSTRACT

Recently, Arabic language has become one of the most used languages in the web. However, the majority of existing solutions to improve web usage do not take into account the characteristics of this language. The process of browsing search results is one of the major problems with traditional web search engines, especially with ambiguous queries.

Using a ranked list as return result of a specific user request is time consuming and the browsing style seems to not be user-friendly. In this paper, we propose to study how to integrate and adapt the Formal Concept Analysis (FCA) as a new system for Arabic Web Search Results Clustering based on their hierarchical structure. The effectiveness of our proposed system is illustrated by an experimental study using Arabic comprehensive set of documents from the Open Directory Project hierarchy as benchmark, where we compare our system with two others: Suffix Tree Clustering (STC) and Lingo. The comparison focuses on the quality of the clustering results and produced label by different systems. It shows that our system outperforms the two others.

1. Introduction

Internet world users-by-language statistics in 2013 show an impressive growth in Arabic speakers on the internet with 135.6 in millions of users.[1] Moreover, the number of Arabic documents available in the Internet is growing at a rapid pace. Therefore, helping Arabic users to find the response to their needs in the web becomes an interesting topic for research. In fact, the process of browsing search results using a ranked list as return result of a specific user request is time consuming and the browsing style seems to not be user-friendly especially with ambiguous query. Generally, most users just view the top results of their query displayed in the first pages and therefore might miss relevant documents. Furthermore, most Arabic documents in the web do not contain any marks of diacritics, consequently widening the gap between user needs and the results presented in the first pages. In such a case, Web Search Results Clustering (WSRC) is of critical importance for online grouping of similar documents to improve and to facilitate browsing web pages in a more compact and thematic form. Many commercial solutions were proposed in the last years such as iBoogie,[2] yippy,[3] Kartoo,[4] Dogpile.[5] However, these solutions were developed especially for languages whose orthography is based on Latin script or use cross-language mapping from Arabic into English to construct different clusters using a variety of clustering algorithms. The challenge is to create a new system for Arabic Web Search Results Clustering. The system would build distinct labeled clusters of web snippets returned by auxiliary search engines, to answer Arabic-speaking users' needs. In this paper, we present a new system of Web Search Results Clustering for Arabic web documents based on the Formal Concept Analysis (FCA) (Wille, 2005). FCA was successfully used as a new way of clustering web search results based on conceptual clustering. It was integrated in many systems to solve the problem of web browsing especially for European languages (Carpineto and Romano, 2004; Cigarrán et al., 2004; Zhang and Feng, 2008). To the best of our knowledge, FCA has never been used for Arabic WSRC to solve the problem of browsing for Arabic internet users. Moreover, Arabic language has its own properties which are very different from European languages, so using any existing European Web Search Results Clustering models directly can negatively impact the cluster results (Moukdad and Large, 2001). Our contribution in this paper is to study how FCA

* Corresponding author.
E-mail address: issam.sah@gmail.com (I. Sahmoudi)

[1] http://www.internetworldstats.com/stats7.htm.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

[2] http://iboogie.com/
[3] http://yippy.com/
[4] http://fr.kartoo.com/
[5] http://www.dogpile.com/

http://dx.doi.org/10.1016/j.jksuci.2016.09.004
1319-1578/

can be applied to the Arabic language and integrated in a new system for Arabic WSRC. The remainder of this paper is organized as follows. In Section 2, we discuss related works. In Section 3, we present the basics of FCA theory. Whereas in Section 4, we suggest integrating FCA in a new scheme in order to get the web increasingly adapted to the Arabic language. Experiments and evaluations are conducted in Section 5. Finally, we provide conclusions and future works in Section 6.

2. Related work

Web Search Results Clustering (WSRC) aims to organize snippets sharing a common topic into the same cluster, and form corresponding labels for description. Recently, it has become one of the central domains of research to solve the web-browsing problem, and many approaches were proposed which can be classified into two categories: Data-Centric and Description-Centric. For more details see (Carpineto et al., 2009) for a survey.

2.1. Data-Centric Approach

The Data-Centric Approach regroups a set of WSRC systems which are based on classical clustering algorithms such as Hierarchical (Kaufman and Rousseeuw, 2005), K-means (Hartigan and Wong, 1979) and Spectral (Planck and Luxburg, 2006), which are applied to group search results and often slightly adapted to produce a meaningful cluster description. This category contains many examples of systems such as Lassi (Maarek et al., 2000), CIIRarchies (Lawrie and Croft, 2003), Armil (Geraci et al., 2006) and Scatter/Gather (Cutting, 1992). Generally, the most critical problem of all approaches in this category is the cluster's label quality. In fact, when a cluster's label quality is given priority in Clustering Search Results, the second category becomes more important in order to produce groups with understandable labels, these labels are not randomly selected, but they have to be related to the topic researched.

2.2. Description-Centric Approach

The first method in this category was proposed by Zamir et al. and was named Grouper (Zamir et al., 1999). It is an online clustering technique based on Suffix Tree Data Structure where search results are clustered and clusters are labeled using the common phrases found by Suffix Tree Data Structure. Suffix Tree Data Structure was adapted in our previous system of WSRC for Arabic language called AWSRC (Sahmoudi and Lachkar, 2013). Another solution in this category, FCA which is a mathematical theory introduced by Rudolf Wille in 1984 has been integrated in many systems of WSRC such as JBreanDead (Cigarrán et al., 2004), Credo (Carpineto and Romano, 2004) and CHC (Zhang and Feng, 2008). Although the FCA has been successfully used as conceptual clustering technique to overcome the problem of WSRC with intentional description of each cluster to make groupings more interpretable, its main drawback is that the concept lattice generated can be unmanageable when applied to large document collections and rich sets of indexing terms (Ch et al., 2015; Cheung and Vogel, 2005; Dias and Vieira, 2010, 2015; Li et al., 2012). In this paper, we study how to integrate FCA in a new system for Arabic WSRC.

3. Formal Concept Analysis Theory

The basic idea of the use of FCA model is to explore the formal context between resulting snippets of ranked items returned by search engines firstly and then construct the concept lattice as new snippets' representation. In this section, we present the Formal Concept Analysis Theory by giving some important definitions and some illustrative examples.

3.1. Formal context (G, M, I)

Formal context (G, M, I) consists of a set of objects G, a set of attributes M, and I is defined by a binary relation between objects G and attributes M in a data set that relates objects with values of the attributes. Table 1 shows an example of formal context.

3.2. Formal concept of formal context (G, M, I)

The formal concept of a formal context (G, M, I) is a set of objects that share similar characteristics. Using the mathematical definition given by Rudolf Wille, the formal concept is defined as a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A = B^I$ and $B = A^I$. A and B are called respectively the extent and the intent of the formal concept (A, B) (Wille, 2005).

Where:

$$A^I = \{m \in M \mid gIm \ \forall g \ \in A\}$$

$$B^I = \{g \in G \mid gIm \ \forall m \in B\}$$

A^I is the derivation operator of A and B^I is the derivation operator of B.

3.3. The concept lattice (G, M, I)

The concept lattice (G, M, I) is an ordered hierarchy of all Formal concepts of the formal context (G, M, I). Many algorithms were proposed to construct the concept lattice from the formal context. They can be classified into two categories:

(a): Algorithms are developed to enhance the performance in generating the set of concepts such as Ganter (2003); (b): Algorithms are developed to enhance performance in building the entire lattice such as Godin's et al. (1995), Bordat (1986) and Nourine and Raynaud (2002). Fig. 1 shows the concept lattice corresponding to the formal context presented in Table 3.

Table 1
Example of formal context.

	Jaguar/جغوار	Car/السيارة	Vehicle/المركبة	Model/نموذج	Sports/الرياضة	Animal/حيوان	Leopard/فهد
G1	1	1	0	0	0	0	0
G2	1	1	0	1	0	0	0
G3	0	1	0	0	1	0	0
G4	1	0	0	0	0	1	0
G5	1	0	1	0	0	0	0
G6	1	1	0	0	0	0	0
G7	0	0	1	1	0	0	0
G8	1	0	0	0	0	1	1
G9	1	1	0	0	1	0	0