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

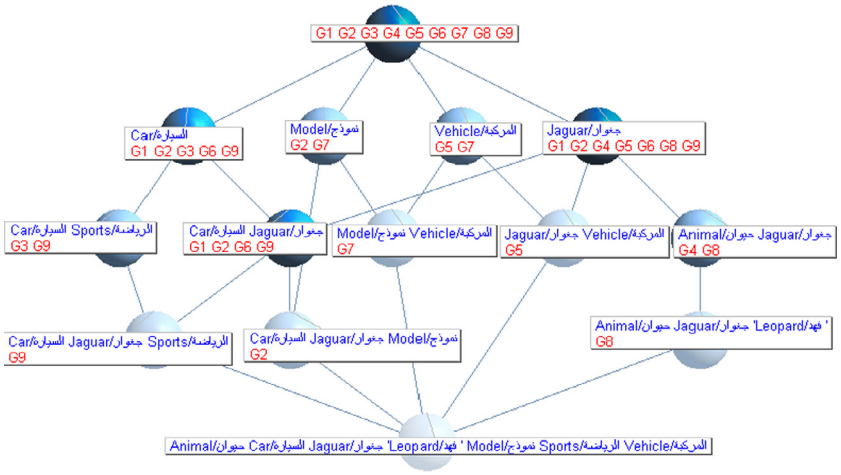

Figure 1. Example of concept lattice generated from formal context in Table 3.

In the following section, we will describe in detail our proposed system for Arabic Web Search Results Clustering based on Formal Concept Analysis. Furthermore, to illustrate the different steps of our proposed system, an example will be presented.

4. Proposed system for Arabic Web Search Results Clustering based FCA theory

The Arabic language is the fourth most spoken language in the internet, and with the number of Arabic documents available on the web increasing at exponential rates, it has become interesting for researchers to propose new information retrieval systems adapted for Arabic users in order to help them find the relevant Arabic web documents. In this section, we present our proposed system for Arabic Web Search Results Clustering to help Arabic users find more pertinent information with their corresponding queries.

4.1. Flowchart

Our proposed new system can be described by a flowchart presented in Fig. 2 and summarized by the following steps:

1. Upload Snippets from Google and Bing.
2. Text Pre-processing.
3. Concept lattice Construction.
4. Clusters Selection.
5. Clusters' Label Generation.

4.2. Formal context construction

The Arabic user specifies his/her query in Arabic language using the web interface. The query is sent to both Google and Bing web search engines using the services offered by the Google API[6] and Bing API.[7] The list of returned results is in the form of Snippets, in order to make our implementation simple and easy, for each Snippet we will associate the following four Tags (ID, URL, Body, and Title) as shown in Fig. 3.

Where:

- ID: Document's Snippet identifier.

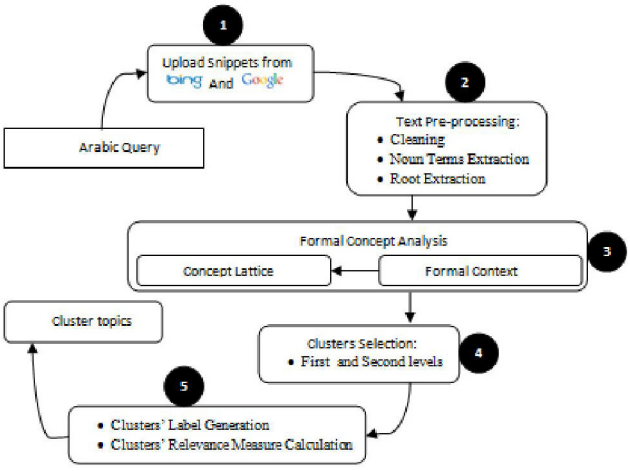

Figure 2. Flowchart of our proposed Arabic system for Web Search Results Clustering based FCA (AWSRC-FCA).

```
<snippet>
  <id>316</id>
  <url>http://www.kuwaityiat.com/</url>
  <body>الموقع الاول لسيدات اعمال الكويت</body>
  <title>كويتيات</title>
</snippet>
```

Figure 3. Example of an Arabic web document's snippet.

- URL: Link to access document content.
- Body: Snippet's Content.
- Title: Page's title.

Each Snippet is cleaned by removing Arabic stop-words, Latin words and special characters like , (/, '\#, \$, etc,...). The fact that treating text mining applications has led us to confirm that the noun terms are the most discriminating terms of document content. Consequently, we propose to add grammatical patterns to select just the noun terms from snippet's content. To extract the

[6] https://developers.google.com/custom-search/docs/start
[7] https://datamarket.azure.com

Table 2
Illustrative example of redundant information removal process.

Object \ Attributes	a	b	c	d
1	1	0	1	1
2	1	0	0	1
3	0	1	0	0
4	1	1	0	1

Redundant information removal

Object \ Attributes	a	b	c
1	1	0	1
2	1	0	0
3	0	1	0
4	1	1	0

noun terms, we used Al-khalil Arabic morphosyntactic analysis system (Boudlal and Lakhouaja, 2010) implemented in Safar platform.[8]

After that, each term will be stemmed to find the corresponding stem. Finally, the obtained stems represent the set of attributes, and Snippets' ID represents the set of objects in the formal context. Table 4 shows an example of obtained formal context using 'SPORT', 'الرياضة' as the user's query. In our case, we define the formal context components as follows:

- Objects: are the Snippets returned from Google and Bing without redundancy, where they are represented by the corresponding ID.
- Attributes: are a set of extracted Roots from each snippet.
- Relation: is a binary relation defined as follows:
  a. True "1": if the word is part of the snippet.
  b. False "0": otherwise.

4.3. Redundant information removal

The main objective of this step is to eliminate the redundant information in the formal context to produce a concept lattice isomorphic to the original. To this end, we propose to adapt the attribute reduction approach. For example, an attribute is redundant if it has exactly the same objects as another. The lower-frequency attribute is then eliminated. Table 2 presents an illustrative example of redundant information removal process.

4.4. Concept lattice Construction and Clusters Selection

The obtained formal context is used to construct the concept lattice. Fig. 4 shows an example of generated concept lattice using 'SPORT', 'الرياضة' as the user's query. In our case, we use the free Java API named ToscanaJ,[9] which integrates Ganter's algorithm (Ganter, 2003) to generate the set of Formal Concepts and the corresponding concept lattice. The latter represents a set of concepts organized in hierarchical structure, where each concept regroups a set of documents (Objects represented by Snippets' IDs in formal context rows) that represent the Extent sharing a set of terms (Attributes represented by terms in formal context columns) that represent the Intent. Note that the concept represents a cluster when using FCA for a clustering process (Carpineto and Romano, 2004; Cigarrán et al., 2004; Zhang and Feng, 2008). In this work, we propose to use the obtained concepts that occur in the first and second levels of the concept lattice Hierarchy as selected clusters. In fact, we select only the first and second levels to get more separated clusters with more descriptive labels. Furthermore, in order to facilitate user's browsing, the obtained clusters must be mapped from the concept lattice to a graphical user interface and presented according to their relevance to the user's query. To this end, we have developed a simple and effective graphical user interface in which the obtained clusters were ranked by taking into account their relevance to the user's query.

Generally, the problem with the cluster relevance ranking is to estimate the relevance of a corresponding concept to a user's query. To overcome this problem, Zhang et al. proposed a new method to construct the two reduced levels hierarchy from the concept lattice. This method is based on two mathematical measures (Zhang and Feng, 2008): the first one is the concept importance measure used to indicate how important the concept is. This measure is relevant to both the number of documents in the extent and the number of descendant concepts of this concept. The second measure is concept similarity, which is based on Jaccard's Similarity Coefficient and it will be used in the merging process to construct a two level hierarchy for the user's browsing.

The system made by Zhang and Feng (2008) is based on the use of all terms extracted from the snippet. Therefore, a reduction process is necessary to reduce a number of not significant generated clusters by filtering or grouping similar concepts. On the other hand, we use stemmed noun terms instead of using all terms without stemming, therefore rendering reduction an unnecessary step. In fact, the number of the noun terms in each snippet is very few. In addition, they are related to the topic of the corresponding document content. However, estimating the relevance of a corresponding concept to a user's query is necessary in order to facilitate the user's browsing process. As we mentioned above, a concept is characterized by two components: the extent and the intent.

Therefore, in this work we propose our new concept relevance measure that takes into consideration the two following components:

- The number of Documents in the Extent.
- The weight of each word in the Intent.

We define our proposed relevance S(Ci) as a measure of Concept Ci as follows:

$$\text{Extent\_Weight} = (|\text{Extent}(Ci)|/\text{Nbr\_Total\_Docs})$$

$$\text{Intent\_Weight} = \sum(\text{TF.IDF}(\text{Intent}(Ci))/|\text{Intent}(Ci)|$$

$$S(Ci) = \text{Extent\_Weight} * \text{Intent\_Weight}$$

Where:
- |Extent(Ci)|: The number of Documents in the Extent.
- Nbr_Total_Docs: The total number of Snippets in the Corpus.
- ∑(TF.IDF(Intent(Ci))/|Intent(Ci)|: The Average of TF.IDF of all words of the Intent in the corresponding concept.

[8] http://sibawayh.emi.ac.ma/safar/download.php
[9] http://toscanaj.sourceforge.net/

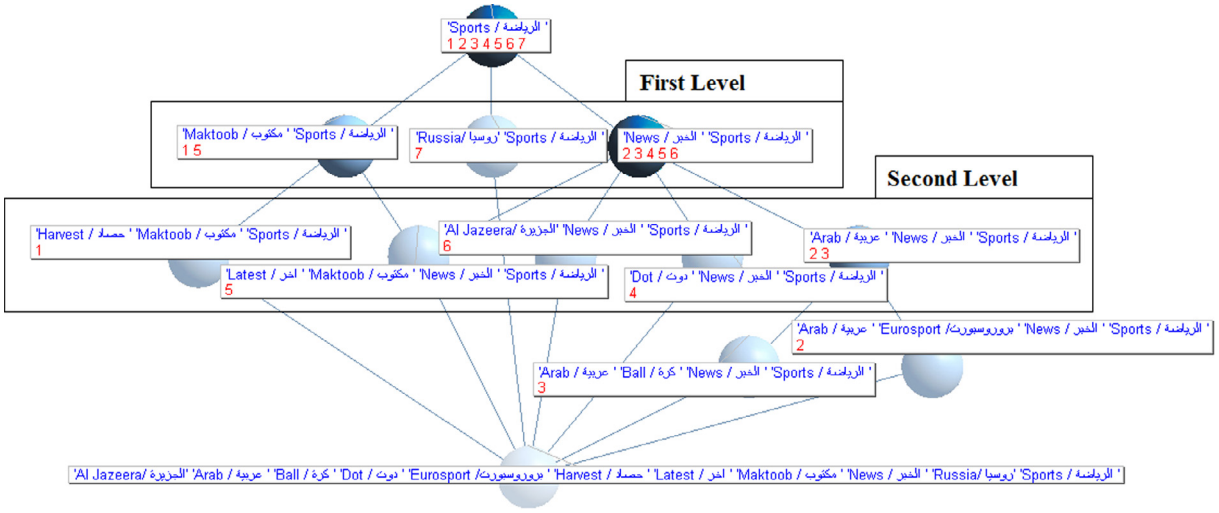

Figure 4. Concept lattice of (Sport, الرياضة) query.

4.5. Cluster's Label Generation

Clusters Label Generation is a crucial step because meaningless or misleading labels may lead users to check the wrong clusters. Moreover, labels should be more comprehensible for users and accurately describe the contents of documents. To this end, we propose to find the original term of each item in the Intent of the corresponding concept, and the Cluster's Label is a comma-separated set of original terms. Then, the user can simply click on the produced Cluster's Label that best describes their specific information needs in the Topic Hierarchy.

4.6. Illustrative example

Next, we present an illustrative example to further explain how our system works. For the sake of simplicity, due to the large number of returned pages, we use seven snippets from the first results page returned by Google and Bing using (Sport, ةضايرلا) as the query (Table 3).

Each snippet is cleaned by removing stop-words. Then, each term in the Snippet is stemmed to find the corresponding Stem. The obtained terms represent the set of attributes in the Formal context as columns and Snippets' IDs represent the rows. Table 4 shows the obtained Formal context of these seven Snippets that correspond to (Sport, ةضايرلا) query.

Table 3
Example of seven page titles of returned results using (Sport, الرياضة) query.

Snippet's ID	Snippets in Arabic and English
1	مكتوب الرياضي – حصاد 2012 الرياضي – Yahoo! Harvest 2012 Sports – Maktoob Sports – Yahoo!
2	اخبار الرياضة ومباريات اليوم من يوروسبورت عربية News, sports and games today Arab Eurosport
3	اخبار الرياضة وكرة القدم – أخبار سكاي نيوز عربية News, sports and football – Arab News Sky News
4	أخبار الرياضة _ رياضة دوت كوم Sports News Sports dot com
5	مكتوب أخبار الرياضة	مكتوب الرياضي	آخر الأحداث الرياضية – Yahoo! Sports News	written Sports	Latest sporting events – Yahoo! Maktoob
6	الجزيرة الرياضية: الأخبار Al Jazeera Sports: News
7	الرياضة – روسيا اليوم Sports – Russia Today

In this step, we eliminate the redundant information in the formal context Table 4. Table 5 presents the formal context after redundant information removal process.

The Formal context presented in Table 5 will be used to construct the Concept Lattice. This is presented in Fig. 4, and will be used for clusters selections, which are presented in a hierarchical structure with different levels. As we observe in Fig. 4, there are three concepts in the First Level which are the more general clusters: (2, 3, 4, 5, 6; رابخا, ةضايرلا /sports, news), (1, 5; بوتكم, ةضايرلا/ sports, maktoob) and (7; ةضايرلا ,ليسور/sport, russia). The user can choose to browse inside any one of the clusters in the first level by clicking on their Labels (رابخا, ةضايرلا /sport, news), (ة, بوتكم ضايرلا /sport, maktoob), (ليسور, ةضايرلا/sport,russia). Then, the user can access more Topics in the second level.

To help the Arabic user find his/her information needs, the clusters will be ranked and displayed according to their weight as illustrated in Fig. 5, which presents three clusters in the first level of the obtained hierarchy corresponding to the sport query. Furthermore, all Clusters' Labels correspond exactly to the initial terms in the Snippet separated by commas. Note that the first cluster, which is labeled by (News, Sport, رابخا, ةضايرلا) is in fact the most relevant to the query (Sport, ةضايرلا).

5. Experiment results and discussion

In this section, we present a comparative study between our proposed system and two others as baseline: STC and Lingo. STC is a classical WSRC based on Suffix Tree Data Structure and Lingo is a well-known WSRC algorithm in the academic field. Both systems are integrated in the carrot2[10] platform, which is an open source search results clustering engine. Note that release 3.2.0 of carrot2 introduces experimental support for clustering content in Arabic. This comparative study is interested in the quality of the clustering results and produced label by different systems.

Open Directory Project (ODP) is the largest, most comprehensive human-edited directory of the web. It's the web, organized. It is constructed and maintained by a passionate, global community of volunteer editors. It is a searchable web-based multilanguage directory consisting of few million web pages preclassified and organized as tree. For Arabic language, the ODP includes 4781 snippets pre-classified into 459 categories by a

[10] http://project.carrot2.org/

Table 4
Formal Context of (Sport, الرياضة) as query.

	Harvest/ حصاد	Sports/ الرياضة	Maktoob/ مكتوب	News/ الخبر	Arab/ عربية	Eurosport/ بوروسبورت	Ball/ كرة	Foot/ القدم	Sky/ سكاي	Dot/ دوت	Latest/ اخر	Events/ الاحداث	Al Jazeera/ الجزيرة	Russia/ روسيا
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	1	1	1	0	0	0	0	0	0	0	0
3	0	1	0	1	1	0	1	1	1	0	0	0	0	0
4	0	1	0	1	0	0	0	0	0	1	0	0	0	0
5	0	1	1	1	0	0	0	0	0	0	1	1	0	0
6	0	1	0	1	0	0	0	0	0	0	0	0	1	0
7	0	1	0	0	0	0	0	0	0	0	0	0	0	1

Table 5
Formal context after redundant information removal process.

	Harvest/ حصاد	Sports/ الرياضة	Maktoob/ مكتوب	News/ الخبر	Arab/ عربية	Eurosport/ بوروسبورت	Ball/ كرة	Dot/ دوت	Latest/ اخر	Al Jazeera/ الجزيرة	Russia/ روسيا
1	1	1	1	0	0	0	0	0	0	0	0
2	0	1	0	1	1	1	0	0	0	0	0
3	0	1	0	1	1	0	1	0	0	0	0
4	0	1	0	1	0	0	0	1	0	0	0
5	0	1	1	1	0	0	0	0	1	0	0
6	0	1	0	1	0	0	0	0	0	1	0
7	0	1	0	0	0	0	0	0	0	0	1

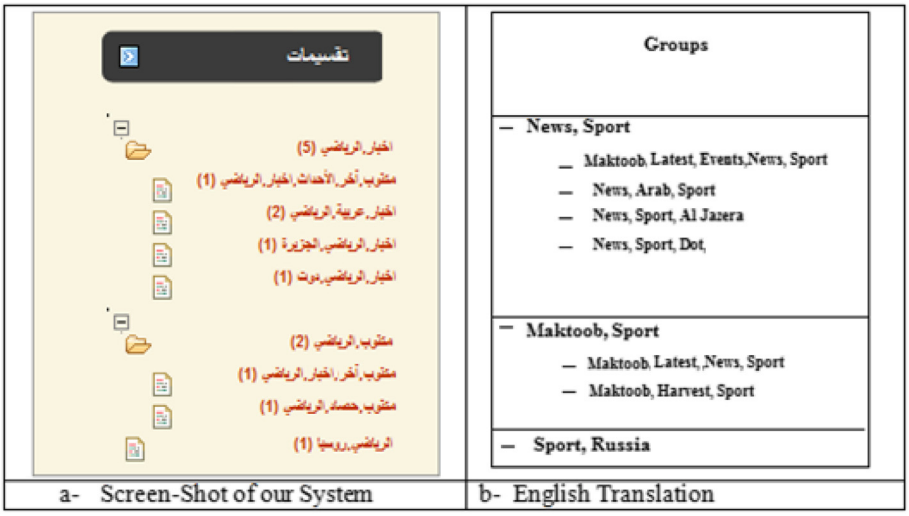

Figure 5. Screen-shot of our system for (Sport, الرياضة) query.

group of human experts. Consequently, the ODP represents a good ground truth for our comparative study.

5.1. Clustering results quality

Generally, the quality of clustering results of any clustering system can be measured by the degree to which it is able to correctly re-classify a set of pre-classified snippets into exactly the same categories without knowing the original category assignment.

The quality of clustering results can be measured by two metrics Normalized Mutual Information (NMI) and Normalized Complementary Entropy (NCE). These metrics are employed by Geraci et al. to compare the effectiveness of different WSRC algorithms (Geraci et al., 2006). For a given a set S of N snippets pre-classified under $C = \{c_1, c_2, \ldots, c_n\}$ of categories and a set $C' = \{c'_1, c'_2, \ldots, c'_m\}$ as the clustering results the NMI and NCE are defined as follows:

$$NMI(C, C') = \frac{2}{\log|C||C'|}\sum_{c\in C}\sum_{c'\in C'}P(c,c')\log\frac{P(c,c')}{P(c)P(c')}$$

Where:

$$P(c) = \frac{|c|}{N},\ P(c') = \frac{|c'|}{N},\ P(c,c') = \frac{|c|\cap|c'|}{N}$$

$$NEC(C, C') = \sum_{i=1}^{m}\frac{|c'_i|}{N'}NCE(C, c'_i)$$

Where:

$$NEC(C, c'_i) = 1 - \frac{2}{\log|C|}\sum_{j=1}^{n} -\frac{P(c_j, c'_i)}{P(c_j)}\log\frac{P(c_j, c'_i)}{P(c_j)}$$

$$N' = \sum_{i=1}^{m}|c'_i|$$

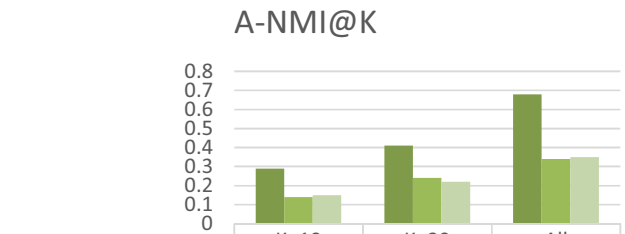

Figure 6. Comparative study: A-NMI@K.

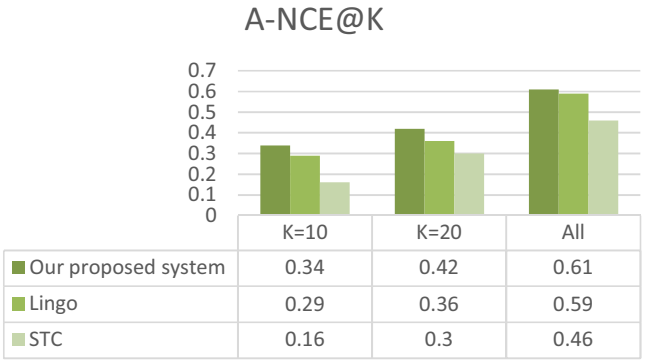

Figure 7. Comparative study: A-NCE@K.

NMI is designed for non-overlapping clustering therefore higher values NMI indicate better clustering quality. NCE range in the interval [0, 1] and is designed for considering overlap, greater value of NCE mean better clustering. Zhang and Feng (2008) shows that these metrics suffer from bias for the following reasons:

1- If the original categories are fixed, the more the clusters generated by a certain WSRC algorithm, the higher the values of NMI and NCE obtained.
2- If the clusters that need to be compared are fixed, the more the groups in the original categories, the higher the values of NMI obtained.
3- When comparing the resulting clusters generated by two different WSRC algorithms with NCE and NMI, the performance may be reverted if using different original categories.

To overcome the above biases of the two metrics, they propose two improved metrics: A-NMI@K and A-NEC@K, where A indicates the average of each result from the used categories and K indicates the number of clusters selected for the experiment. The use of the improved metrics considers the variation of the experimental results when changing the category, and can provide a global idea about the performance of the systems used for the experiment. In this comparative study, we use only K = 10, K = 20 and K = all because we consider that 10 is the minimal number of clusters visited by a user and 20 is the maximal one. Also, the results remain the same for both K = 5 and K = 15.

Fig. 6 presents the A-NMI@10, A-NMI@20 and A-NMI@ALL for non-overlapping clustering quality measures of the three systems: Our system, STC and Lingo. It is clear that our system outperforms the two other ones and provides a twice better improvement regard to the two others systems. Fig. 7 presents the A-NCE@10, ANCE@20 and A-NCE@ALL for overlap clustering quality measure, and it shows that our system outperforms the two others.

5.2. Cluster's label quality

The main goal in this subsection is to compare the Labeling quality of a cluster for the three systems. In fact, looking into WSRC state of the art reveals a serious issue in cluster label quality evaluation. Generally, using human expert evaluation is not always possible, due to the lack of human resources. Furthermore, human experts may not be able to evaluate several thousands of queries especially with an online system such as WSRC. Therefore, we propose to use only two queries to get an idea about the quality of the Cluster label for each system. The first one is (Commerce, ةراجت) and the second one is (Education, ميلعتلا).

Figs. 8 and 9 show the labels of the 10 clusters produced by our System, Lingo and STC for the two queries. According to our team which we consider as experts, the labels produced by the three

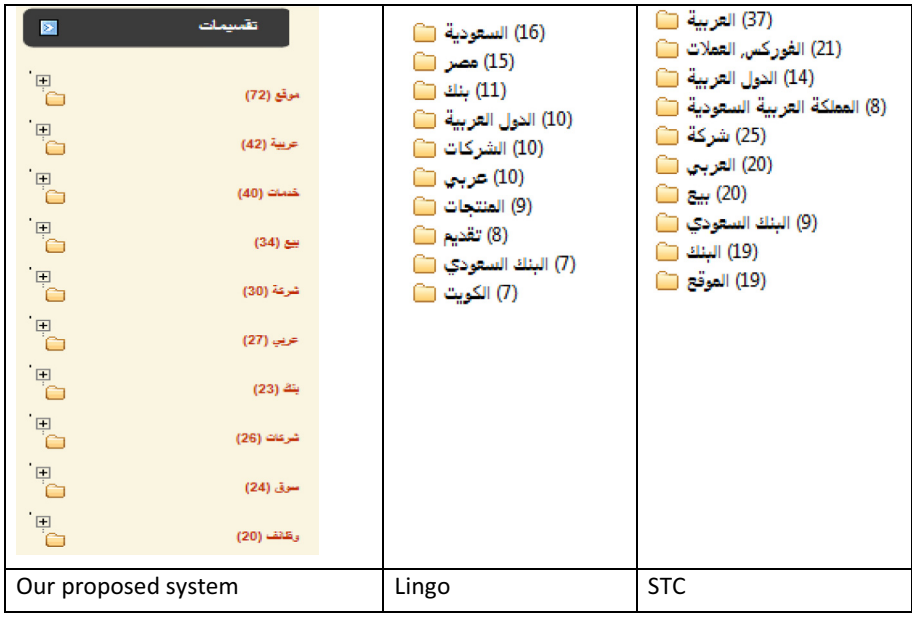

Figure 8. Cluster's results of (Commerce, تجارة) query.

Figure 9. Cluster's results of (Education, التعليم) query.

systems are readable and informative. However, each cluster generated by Lingo or STC only consists of a few documents, the number indicated by the number followed each label, which means that many documents are not grouped into the 10 clusters of both Lingo and STC. On the other hand, the labels produced by our system is not based on phrases but based on keywords. Each label is composed of one or more keywords. It provides the discovery of causal association between words that hold in the set of results. Generally, the results presented in both the queries have proven the effectiveness of our system compared to the other.

6. Conclusion

Browsing Search Results is one of the major problems with traditional web search engines (Google, Yahoo and Bing) for English, European, and any other language in general, and for the Arabic language in particular. Organizing Arabic web search results into clusters facilitates browsing the web for Arabic users. In this paper, we have proposed the use of Formal Concept Analysis in a new system of WSRC for the Arabic language. The proposed system automatically clusters the web search results into high quality clusters with a hierarchical structure, and provides descriptive cluster labels. A series of experiments was conducted: both subjective and objective evaluations were presented using Google search and Bing search APIs. The results obtained were very encouraging and illustrate the efficiency of our proposed system. In future works, we believe that it could be possible to improve the performance of our proposed system by integrating some external knowledge resources such as Arabic Word-Net and Arabic Wikipedia.

References

Bordat, J., 1986. Calcul pratique du treillis de Galois d'une correspondance. Math. Sci. Hum. Math. Soc. Sci. 96, 31–47.

Boudlal, A., Lakhouaja, A., 2010. Alkhalil morpho SYS1: A morphosyntactic analysis system for Arabic texts. Int. Arab Conf. Inf. Technol., 1–6

Carpineto, C., Romano, G., 2004. Exploiting the potential of concept lattices for information retrieval with CREDO. J. Univers. Comput. Sci. 10, 985–1013.

Carpineto, C., Osiński, S., Romano, G., Weiss, D., 2009. A survey of Web clustering engines. ACM Comput. Surv. 41, 1–38. http://dx.doi.org/10.1145/1541880.1541884.

Ch, A.K., Dias, S.M., Vieira, N.J., 2015. Knowledge reduction in formal contexts using non-negative matrix factorization. Math. Comput. Simul. 109, 46–63. http://dx.doi.org/10.1016/j.matcom.2014.08.004.

Cheung, K.S.K., Vogel, D., 2005. Complexity reduction in lattice-based information retrieval. Inf. Retr. Boston. 8, 285–299. http://dx.doi.org/10.1007/s10791-005-5663-y.

Cigarrán, J.M., Gonzalo, J., Peñas, A., Verdejo, F., 2004. Browsing search results via formal concept analysis: automatic selection of attributes. Concept Lattices 2961, 201–202. http://dx.doi.org/10.1007/b95548.

Cutting, D.R., 1992. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections 1 Introduction 2 Scatter/Gather Browsing.

Dias, S.M., Vieira, N., 2010. Reducing the size of concept lattices: the JBOS approach. In: Cla 2010, pp. 80–91.

Dias, S.M., Vieira, N.J., 2015. Concept lattices reduction: definition, analysis and classification. Expert Syst. Appl. 42, 7084–7097. http://dx.doi.org/10.1016/j.eswa.2015.04.044.

Ganter, B., 2003. Ch1 & Ch2: Contexts, concepts, and concept lattices. Form. Concept Anal. Methods Appl. Comput. Sci.

Geraci, F., Pellegrini, M., Maggini, M., Sebastiani, F., 2006. Cluster generation and cluster labelling for Web snippets: a fast and accurate hierarchical solution. String Process. Inf. Retr. 13, 25–36. http://dx.doi.org/10.1007/BF02959914.

Godin, R., Missaoui, R., Alaoui, H., 1995. Incremental concept formation algorithms based on Galois (concept) lattices. Comput. Intell. 11, 246–267. http://dx.doi.org/10.1111/j.1467-8640.1995.tb00031.x.

Hartigan, J.A., Wong, M.A., 1979. A K-means clustering algorithm. J. R. Stat. Soc. 28, 100–108. http://dx.doi.org/10.2307/2346830.

Kaufman, L., Rousseeuw, P.J., 2005. Finding groups in data: an introduction to cluster analysis. Intensive Care Med. 33, 368. http://dx.doi.org/10.1007/s00134-006-0431-z.

Lawrie, D.J., Croft, W.B., 2003. Generating hierarchical summaries for web searches. In: Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. informaion Retr. – SIGIR '03 457. http://dx.doi.org/10.1145/860435.860549.

Li, J., Mei, C., Lv, Y., 2012. Knowledge reduction in real decision formal contexts. Inf. Sci. (Ny) 189, 191–207. http://dx.doi.org/10.1016/j.ins.2011.11.041.

Maarek, Y.S., Fagin, R., Ben-shaul, I.Z., Pelleg, D., 2000. Ephemeral document clustering for web applications. IBM Res. Rep. RJ 10186, 1–26. http://dx.doi.org/10.1007/s00417-013-2383-7.

Moukdad, H., Large, A., 2001. Information retrieval from full-text Arabic databases: Can search engines designed for English do the job? Libri 51, 63–74. http://dx.doi.org/10.1515/LIBR.2001.63.

Nourine, L., Raynaud, O., 2002. A fast incremental algorithm for building lattices. J. Exp. Theor. Artif. Intell. 14, 217–227. http://dx.doi.org/10.1080/09528130210164152.

Planck, M., Luxburg, U. Von, 2006. A tutorial on spectral clustering a tutorial on spectral clustering. Stat. Comput. 17, 395–416. http://dx.doi.org/10.1007/s11222-007-9033-z.

Sahmoudi, I., Lachkar, A., 2013. Clustering web search results for effective Arabic language browsing. Int. J. Nat. Lang. Comput. 2, 31–43. http://dx.doi.org/10.5121/ijnlc.2013.2202.

Wille, R., 2005. Formal concept analysis as mathematical theory of concepts and concept hierarchies. Form. Concept Anal. 1–33. http://dx.doi.org/10.1007/11528784_1.

Zamir, O., Etzioni, O., 1999. Grouper: A dynamic clustering interface to Web search results. In: Proc. WWW8.

Zhang, Y., Feng, B., 2008. Clustering search results based on formal concept analysis. Inf. Technol. J. http://dx.doi.org/10.1109/FSKD.2008.140.