Creating Facets Hierarchy for Unstructured Arabic Documents

Khaled Nagi and Dalia Halim
Dept. of Computer and Systems Engineering, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Keywords: Faceted Search, Arabic Content on the Internet, Indexing Arabic Content.

Abstract: Faceted search is becoming the standard searching method on modern web sites. To implement a faceted search system, a well defined metadata structure for the searched items must exist. Unfortunately, online text documents are simple plain text, usually without any metadata to describe their content. Taking advantage of external lexical hierarchies, a variety of methods for extracting *plain* and *hierarchical* facets from textual content are recently introduced. Meanwhile, the size of Arabic documents that can be accessed online is increasing every day. However, the Arabic language is not as established as the English language on the web. In our work, we introduce a faceted search system for unstructured Arabic text. Since the maturity of Arabic processing tools is not as high as the English ones, we try two methods for building the facets hierarchy for the Arabic terms. We then combine these methods into a hybrid one to get the best out of both approaches. We assess the three methods using our prototype by searching in real-life articles extracted from *two* sources: the BBC Arabic edition website and the Arab Sciencepedia Website.

1 INTRODUCTION

Faceted search is simply looking at a search result-set from many different perspectives at the same time. The first perspective is the whole result-set. Then each facet describes a subset of the result-set from different perspectives. Facets can be either *flat* or *hierarchical*. A good hierarchical facet structure is *not deep*. It has a *reasonable number of siblings* for each node. The *expanded path from root to leaf generally fits on one screen* without scrolling down.

In order to implement a faceted search system, a well defined metadata structure for the searched items must exist. Structured data such as product description in an online shop are very good candidates for being used in a faceted search environment. Introducing faceted search for unstructured data, e.g., text documents, has another degree of complexity. Text documents are plain text usually without any metadata to describe their content. Taking advantage of *term extraction tools* and *external lexical hierarchies*, a variety of methods for extracting hierarchical facets from plain text are recently introduced.

However, term extraction tools and external lexical hierarchies are very language specific. The Arabic language is not as established as the English language on the web. Meanwhile, the size of Arabic documents that can be accessed online is increasing every day. There is few to no research done in choosing the correct constellation of emerging Arabic-specific term extraction tools and external lexical hierarchies to create facets.

In our work, we introduce an automatically generated faceted search system for unstructured Arabic text. The developed prototype receives the user query and returns a Google-like result-set. Additionally, it returns a set of hierarchical facet terms to help the user filter the returned result-set to navigate to the required document. For the facet creation process, we rely on LingPipe Named Entity Recognizer (LingPipe, n.d.). In order to create the facets hierarchy from the results terms, we use *two* methods. The first method uses the Arabic Wikipedia hierarchy (Arabic Wikipedia Categorization, n.d.) to build the facet hierarchy for the result-set. The second method uses an English tool to alleviate for relatively poor Arabic corpus in online tools. The facets are translated to English and the hierarchy is built using the English WordNet (WordNet, n.d.) IS-A hypernym structure. Then the whole facets hierarchy is translated back to Arabic. This workaround is used till an Arabic version of Wordnet with a full-fledged corpus is developed by the linguistic researchers (Arabic WordNet, n.d.). In a later stage of our work, we investigate a combination of both methods by

Nagi K. and Halim D..
Creating Facets Hierarchy for Unstructured Arabic Documents.
DOI: 10.5220/0004621201090119
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2013), pages 109-119
ISBN: 978-989-8565-81-5

intelligently merging the best subsets of each of the previously built hierarchies.

We assess the three methods that we built in our prototype using two sets of articles: one extracted from the BBC Arabic edition website (BBC Arabic, n.d.) and the other from the Arab Sciencepedia website (ArabSciencepedia, n.d.). We compare the resulting facet hierarchies in terms of relevance, the shape of the facet hierarchy, and the creation time.

We are aware that the comparison is the result of a clear competition between a wiki-based approach – and hence organically evolving – with a huge community of contributors versus a well structured approach driven by linguistic science. In our work, we train our system using existing corpus in some domain knowledge. So, it is a nice scientific challenge to assess the proposed facet creation mechanisms and studying the effect of the available training corpus and its linguistic structure on the quality of the facets.

The rest of the paper is organized as follows. Section 2 provides an overview of related work. Our proposed system is presented in Section 3. In Section 4, the developed prototype is explained. Section 5 contains an assessment of the generated facet hierarchies while Section 6 concludes the paper.

2 RELATED WORK

2.1 Evolution of Information Retrieval to Faceted Search

Information Retrieval is finding material - usually documents - of an unstructured nature - usually text- that satisfies an information need from within large collections (Manning et al., 2009). The earliest information retrieval systems use a *set retrieval* model (Manning et al., 2009). Set retrieval systems return unordered document sets. The query expressions require Boolean operations and the exact query keywords are supposed to exist or never exist in the searched documents.

Seeking an alternative to the set retrieval model and the complicated query form, the *ranked retrieval* model was introduced. A free-text query is submitted and the system returns a large result-set of documents, ordered by their relevance to the user query. One famous example of the ranked retrieval model is the Vector Space Model (Manning et al., 2009).

The *directory navigation* model was introduced next. This model offers users an advantage over free text search by organizing content in taxonomy and providing users with guidance toward interesting subsets of a document collection. The web directory that Yahoo! built in the mid-1990s (YahooHistory, n.d.) is a nice example. Yet, taxonomies have a main problem. Information seekers need to discover a path to a piece of information in the same way that the taxonomist created it, as each piece of information is categorized under one taxonomy leaf. Actually, if a group of taxonomists is building taxonomy for several terms, they would disagree among themselves on how information should be organized. This shortcoming in the directory navigation is known as the vocabulary problem (Tunkelang, 2009).

Faceted search appeared to be a convenient way to solve the vocabulary problem. Facets refer to categories, which are used to characterize items in a collection (Hearst, 2008). In a collection of searched items, a faceted search engine associates each item with all facet labels that most describe it from the facets collection. This makes each item accessible from many facets paths thus decreasing the probability of falling into the vocabulary problem. Faceted search assumes that a collection of documents is organized based on a well-defined faceted classification system or that the underlying data is saved in a database with a predefined metadata structure. Unfortunately, this is not the case for collections of *unstructured text* documents.

2.2 Facet Creation for Unstructured Documents

Data clustering is a class of solutions for creating subject hierarchies for unstructured documents. The advantage of clustering is that it is fully automated, but its main disadvantages are their lack of predictability and the difficulty of labeling the groups. Some automatic clustering techniques generate clusters that are typically labeled using a set of keywords (Hearst and Pedersen, 1996). Such set of titles gives good indication of a collection of documents contents, but its presentation is very hard to be used by the end users in a navigational interface.

Other technique for enriching unstructured text with metadata is the subsumption algorithm (Mark and Croft, 1999). For two terms x and y, x is said to subsume y if: $P(x|y) \geq 0.8$, $P(y|x) < 1$ (Stoica et al., 2007); which means, x subsumes y and is a parent of y, if the documents which contain y, are a subset of the documents which contain x.

Another class of solutions makes use of existing *lexical hierarchies* to build facets/categories hierarchies. Examples of existing lexical hierarchies are WordNet hypernym and hyponym structures

(WordNet, n.d.) and the Wikipedia categorization structure (Wikipedia Categorization, n.d.). Some of the previous works in the faceted search interface creation for unstructured documents include Castanet (Stoica and Hearst, 2004) and (Stoica et al., 2007), facets for text database (Dakka and Ipeirotis, 2008), Facetedpedia (Yan et al., 2010), and facets for Korean blog-posts (Lim et al., 2011) as an example of Facet creation in non-latin languages.

2.3 Facet Creation Research Projects

The *Castanet* algorithm - (Stoica and Hearst, 2004) and (Stoica et al., 2007) - assumes that there is a text description associated with each item in the collection. E.g., if the collection is a set of images, each image has an *unstructured* text document describing the image. The textual descriptions are used to build the facet hierarchies and then to assign documents to facets.

The *target terms set* is a subset of the terms that best describes the set of documents. Their *selection criterion* is the term distribution. Terms with term distribution greater than or equal to a specified threshold are retained as the target terms. The target terms set is divided into two categories:

- Ambiguous terms: having more than one meaning in the English WordNet.
- Un-ambiguous terms: with only one meaning in the English WordNet.

The core hierarchy is first built for the unambiguous terms using the WordNet IS-A hypernym structure (WordNet, n.d.). Then, the ambiguous terms are checked against the WordNet Domains (WordNet Domains, n.d.); which is a tool assigning domains to each WordNet synonym set. The tool counts occurrences of each domain for unambiguous target terms, resulting in a list of the most represented domains in the set of documents. If the ambiguous term has only one common domain for all its senses in WordNet, it is considered unambiguous. The core hierarchy is next augmented by the unambiguated terms IS-A hypernym paths. A refinement step is next done by compressing the final hierarchy. Nodes with number of children less than a threshold and nodes whose names appear in their parents' node name are eliminated. Finally, in order to create a set of sub-hierarchies, the top levels (e.g., 4 levels) are pruned. Thus, the final facets hierarchy is created.

The algorithm for creating facets for text databases is presented in (Dakka and Ipeirotis, 2008). It is built on the observation: *The terms for the useful facets do not usually appear in the documents contents.* Thus, the target terms list is created using two sets of terms.

- The *first* set includes the significant terms extracted from the document body text using extraction tools.
- The *second* set is created by expanding the first set with other relevant terms using WordNet hypernym, Wikipedia contents, and the terms that tend to co-occur with the first set of terms when queried against the Google search engine.

The *term* frequency is used in the original and the expanded terms set to identify the final candidate facets. Infrequent terms from the first terms set with the frequent terms from the second expanded terms set form together the final set of facets.

Facetedpedia (Yan et al., 2010) is a project that dynamically generates a query-dependent faceted interface for Wikipedia searched articles. The next definitions build the main concepts used in the algorithm.

- *Target Articles*: are the articles in the returned result-set of the user query.
- *Attribute Articles*: each Wikipedia article that is hyperlinked by a target article.
- *Category Hierarchy*: Wikipedia category hierarchy is a connected, rooted directed acyclic graph.

One large hierarchy is built for the target articles as follows. Each target article is connected to all its attributes articles. Then, a category hierarchy is built for each attribute article. Hierarchies for all attribute articles are merged until we find one common root category. Then, the most appropriate set of subcategories is chosen from within the built hierarchy using a cost measurement and a similarity measurement developed by the author.

The work in (Lim et al., 2011) is an example of facet creation for non-Latin languages. Non-Latin languages such as Korean or Arabic do not have powerful linguistic tools when compared to the English WordNet. Workarounds are found by researchers working with these languages. In (Lim et al., 2011), the system generates flat facets interface for Korean blog-posts. Given a search query keyword, blog posts are searched using the search engine "Naver Open API" for Korean (Naver, n.d.). For the initial keyword, a set of blog posts is constructed, where each post with its body text are successfully extracted from the blog post. The facet generation process is done in five steps. The system collects Wikipedia articles that include the user query and extracts the titles of these Wikipedia pages to use them as facets candidates for the blog-posts. After constructing the candidate facets terms set, only the

facets terms that appear more frequently are retained as actual facets terms. Given a Wikipedia entry, terms that are closely related to the term are automatically extracted. Typical closely related terms include bold-faced terms, anchor texts of hyperlinks and the title of a redirect, etc. Then, from each blog post, a term frequency vector is generated. The similarity of a Wikipedia entry and a blog post is defined as the inner product of the inverse document frequency vector of the Wikipedia entry and the term frequency vector. To each blog post, the system assigns a facet that maximizes the similarity of the facet and the blog post. One important Shortcoming of this work is that *no hierarchy* is generated.

3 PROPOSED SYSTEM

While Arabic content increases day per day on the Internet, tools accessing and manipulating Arabic contents are not increasing at the same pace. Thus implementing a hierarchical facets search interface for Arabic unstructured document motivates our idea of trying two different ways in implementing the hierarchy: one using Arabic tools and the other using English tools.

In our implemented system, a hierarchical faceted interface is created in two main steps: *the facets extraction* phase and *the facets hierarchy creation* phase. The facets creation phase results in a set of target terms that most describe the list of searched documents. This target terms set is an input to the next phase, where two parallel processes use this set. The first one produces a facets hierarchy using Arabic tools. The other one translates the set of terms into English, builds the facets hierarchy using English tools, and the hierarchy is translated back to Arabic.

3.1 Facets Extraction Phase

The user query is sent to Google search engine using Google custom search API (Google API, n.d.). The resulting set of documents is processed in order to extract the most significant terms. The significant terms extraction is done using the LingPipe tool (LingPipe, n.d.).

The LingPipe tool provides three different types of significant terms extraction:

- Rule based extraction model,
- Dictionary based model, and
- Statistical model.

The *rule based* extraction model needs a regular expression as input, and then for each processed document, all matches for the regular expression are extracted by LingPipe. The *dictionary based* model uses a list of terms (single or multi-words) as dictionary. Each match from the dictionary in the processed document is considered as significant term. Finally, the *statistical extraction* model requires a training data in the CoNLL format. The training data should match the documents contents on which the extraction will take place. *In our system, we use all 3 models.*

For every regular expression, term in a dictionary, or even term within a training corpus, a *Tag* value is provided. This tag value gives a hint about the term which helps in the term manipulation in the next step of the algorithm.

Since dates play as major role in creating facets, they are handled separately. For example, we handle an extended format found in some Arabic documents, where the date includes the month name in the old Syriac language then its equivalent in the Gregorian calendar ended by the year, e.g., " حزيران ٢٠١٢ (يونيو)" for "June 2012". The rule based extraction model is used in our algorithm for extracting the date from Arabic documents, where a regular expression (in the *rule* based model) for the date format is formulated and fed to the LingPipe extractor. A date extracted using this regular expression is associated with the "Date" tag and is modeled in the standard ISO-date format.

Today the Arabic Wikipedia includes more than 200,000 articles (Arabic Wikipedia, n.d.) covering a very wide range of terms and expressions for Arabic and international concepts. These concepts include historical & contemporary events, person names, locations, scientific terms and many other concepts. Therefore, the Arabic Wikipedia articles titles are used as dictionary entries for the LingPipe *dictionary* based extraction model. A tag "Wiki" is associated with each term during the dictionary build process.

The LingPipe *statistical extraction* model needs a training data, covering the same domains of the documents being processed. Training data is not easily available for all domains. Therefore, the use of such model must not be a key element in the algorithm, but as its usage gives richer results. Our implementation is adjustable to include such models when available, and skip them otherwise. Two data sets are used for testing, the first is news articles for the Arabic BBC website, and the second is scientific articles from the Arab Sciencepedia website. Benajiba's Arabic Named Entity Recognition (BinAjiba, n.d.) training corpus is a CoNLL formatted Arabic training data, which includes terms with location

"Loc", organization "Org" and person "Per" tags values. This corpus is created from several news agencies articles, which makes it appropriate for training LingPipe for the use in the algorithm while searching the Arabic BBC website. The scientific articles facets are created without the statistical extraction model. If statistical extraction model is used, a set of predefined facets is created. This set of facets is the mapping of the tags values appearing in the training data set. For the Arabic news articles training corpus, the predefined facets are "Locations", "Organizations" and "Persons".

3.1.1 Term Reduction Steps

On either data set, the terms extraction process generates more than 6,000 terms for an average of 50 searched documents. Duplicates and synonyms terms are merged simplifying the list to about 3,000 terms. The synonym terms detection is done using the Wikipedia redirect feature (Wikipedia Redirect, n.d.). If two extracted terms are marked to be redirected to one another or to one common article title, then these terms are synonyms and are unified into one term. This terms list is further reduced by omitting terms with low term distribution. The final target terms list contains about 300 terms. This list is the input for the facets hierarchy creation phase. In Figure 1, the term reduction steps are illustrated.

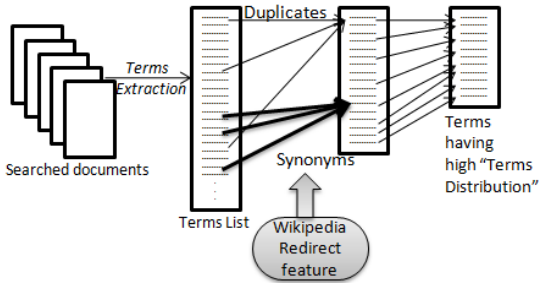

Figure 1: Term reduction steps.

3.2 Facets Hierarchy Creation Phase

During our research, we generate two different facets hierarchies by using two external hierarchies: the *Arabic Wikipedia categorization hierarchy* (Arabic Wikipedia Categorization, n.d.) and the *English WordNet IS-A hypernym* structure (WordNet, n.d.). The usage of one Arabic hierarchy and another English hierarchy is done for testing the feasibility of creating a facets hierarchy for a set of documents with a specified language using an external hierarchy from another foreign language. For the latter case, we use the English WordNet IS-A hypernym structure since it showed relevant success for building facets hierarchies for English documents (Stoica et al., 2007).

3.2.1 Using Arabic Wikipedia Categorization Structure

The documents are classified under a huge categorization graph called محتويات ويكيبيديا (*Wikipedia contents*) (Arabic Wikipedia Categorization, n.d.). The Wikipedia sub-category of interest for our work is التصنيف_الرئيسي (*The Main Categorization*). The main categorization hierarchy is used to build the facets hierarchy for the target terms extracted from the document body text.

Each target term is checked against the Wikipedia articles titles. Once found, the Wikipedia hierarchy for the term is built up to three levels depth only. The number of three levels is chosen as the Wikipedia categorization is very deep, thus a *shortcut* step is taken by building only three levels of the hierarchy.

If a training corpus covering the searched data domain exists, then the statistical terms extraction model is used. The term tag for each extracted term is checked against the predefined facets terms and the term is classified under its corresponding predefined facet if applicable. The predefined facets in the Wikipedia hierarchy are always plain facets. Merging the predefined facets with their corresponding facets in the Wikipedia categorization is not applicable since the categorizations do not match. For example, a name of a country in the Wikipedia categorization can be an upper node of the name of a scientist born in this country. In this case, a predefined facet, such as *location*, would be a parent node for the *country name* followed by a *person name*, which is not applicable.

As the Wikipedia categorization contains some cycles, a depth first traversal is done in order to remove any existing loops and convert the final graph into a hierarchy. Finally, a tree minimization process takes place to enhance the final structure of the hierarchy presented to the end user. The tree minimization is done using three different methods as follows.

- Removing single child vertex: as illustrated in Figure 2, any hierarchy vertex with only one child vertex is omitted and its direct child takes the place of its parent.

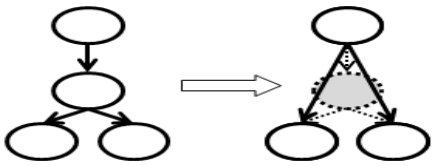

Figure 2: Removing single child vertex.

- Removing redundant sub-trees: as illustrated in Figure 3, sub-trees occurring more than once in different depth levels are omitted from the deeper level.

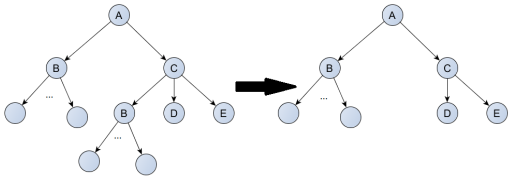

Figure 3: Removing redundant sub-trees.

- Removing facets with little information: The children of the root node are the facets to be presented to the end user. If a facet is only leading to less than three articles, this facet delivers low quality information. It is thus removed from the root's children.

3.2.2 Using WordNet IS-A Hypernym Structure

This phase undergoes five steps as illustrated in Figure 4:

- Term translation,
- Hierarchy construction,
- Term disambiguation,
- Hierarchy minimization, and
- Hierarchy translation back to Arabic.

During the *term translation* step, the Bing translator API (Bing Translator, n.d.) is used. Bing uses the statistical machine translation (SMT) paradigm (Tripathi and Sarkhel, 2010), which means that a document is translated according to the probability distribution $p(e|f)$ that a string e in the target language is the translation of a string f in the source language.

In order to avoid the double translation, from Arabic to English then back to Arabic, the Arabic terms are temporary stored in memory with their translated English meanings.

During the *hierarchy construction* step each term tag is checked against the predefined facets if the LingPipe statistical terms extraction model is used. If the tag matches any of them, the term is classified under its corresponding facet. In the WordNet IS-A structure, the predefined facets are merged with their corresponding WordNet entities.

The WordNet Domains tool (WordNet Domains, n.d.) is used to disambiguate terms that have several senses within WordNet. A "Document-Domains" index containing each document with all domains appearing within the document is created in parallel

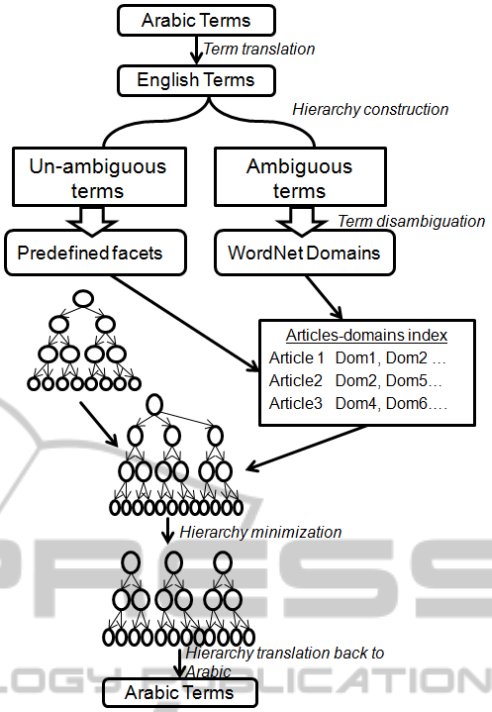

Figure 4: Five steps hierarchy construction using WordNet IS-A.

with the next step. Each translated term is checked against the English WordNet corpus, three different results are possible:

- The term is not found within the English WordNet: the term is dropped unless it falls under a predefined facet.
- The term has only one meaning within WordNet: the IS-A hierarchy is built for this term up to the WordNet root. Then, WordNet domains are checked for this term. Then all term domains are fetched and inserted in the Document-Domains index for all documents containing the term.
- The term has several meanings within WordNet: e.g., the word capital in the English WordNet can be found with the meaning *assets*, *upper-case* or *city*. In this case, the term is associated with an ambiguous term list in order to be disambiguated in the next step.

By now, the hierarchy is built for the un-ambiguous terms, and the Documents-Domains index is created. During the *term disambiguation* step, the following is done for each ambiguous term:

- If all the term senses belong to one domain from the WordNet Domains tool, the first sense is chosen.
- Otherwise, for each document containing the term, a sense with a domain already appearing

in the document is chosen. E.g., the term "Apple" can have both "Food" and "Technology" domains; meanwhile a document that talks about nutrition most probably contains the domain "Food".

- If none of the above is valid, we use the first sense in WordNet since it is usually the most common one. (Stoica and Hearst, 2006).

After the disambiguation step, the built hierarchy is enlarged with the un-ambiguated terms. For each un-ambiguated term, the WordNet IS-A hypernym path for the term is added to the unambiguous hierarchy.

During the *hierarchy minimization* step the system converts the constructed hierarchy into moderate facets sub-hierarchies; each with a reasonable depth and breadth for each level. This is done by removing single child vertex, removing redundant sub-tree and removing facets with little information as described for the Wikipedia hierarchy. Additionally, the first three levels are removed in the compression step since the highest level in WordNet tree contains very abstract terms; such as objects, entities, etc.; which are not very useful as facets.

In the last step, only the inner nodes in the hierarchy are *translated from English to Arabic*.

3.2.3 Using a Hybrid Hierarchy

A normal extension of the work is creating the facets by combining both approaches mentioned in Sections 3.2.1 and 3.2.2. Each hierarchy is created in a separate thread and then a merging step is performed as soon as both hierarchies are ready. During the merging step, the predefined facets from the WordNet hierarchy are merged with the remaining facets from the Wikipedia hierarchy. This step takes place only when a training corpus is available for the use by LingPipe terms extractor for the statistical terms extraction model.

4 THE PROTOTYPE

We develop a hierarchical faceted search system for Arabic documents. The user enters the query and waits for the result set to be returned in a faceted interface as illustrated in Figure 5.

The system is implemented in Java as a web-application under Tomcat server. MySQL is used as database management system. The database is initialized with all data from the Arabic Wikipedia and the English WordNet after preparation and tuning for performance.

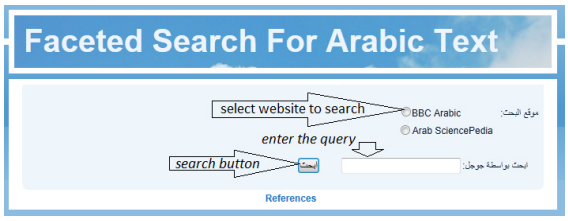

Figure 5: Query entry screen.

When the user enters a search query, this query is sent to the Google search engine using the Google custom search API (Google API, n.d.). The Google custom search API returns a JSON (JSON, n.d.) encoded result-set; which is converted into java objects using the Gson API (GSON, n.d.). Each result item is a complete web page. The Jsoup API (JSOUP, n.d.) is used to extract the main text document from the whole HTML file. The most significant terms appearing in the searched documents are then extracted using the LingPipe tool (LingPipe, n.d.).

Facets created using the three methods are displayed on separate tabs as illustrated in Figure 6.

Figure 6: Result-set with facets.

5 THE VALIDATION

The system prototype is assessed by searching in real-life articles extracted from two sources: the BBC Arabic edition website and the Arab Sciencepedia website. The main differences between both sources are: *the diversity of news versus the concise nature of the scientific articles* and *the availability of training corpus for the new articles*. We compare the resulting facet hierarchies in terms of relevance, structure of the hierarchies as per number of facets and the depth of the hierarchy, and the creation time.

5.1 Relevance of Terms, Articles, and Facets

5.1.1 Terms Coverage

The set of candidate terms is a subset of the set of terms extracted from the text articles. These terms are used in building the facets hierarchies. The final set of terms that appears in the hierarchies is usually only a subset of candidate terms. It depends on the ability of the external resource –Wikipedia and/or WordNet- to recognize the terms and on the hierarchy compression step. We define the terms coverage to be the number of terms in the facets hierarchy divided by the number of candidate terms. Any loss in the terms coverage is not preferred.

As shown in Figure 7, Wikipedia hierarchy covers 80% of the target terms while the loss of terms in the WordNet hierarchy achieves more than 50%. The bad performance of WordNet for the scientific articles is attributed to the lack of training corpus.

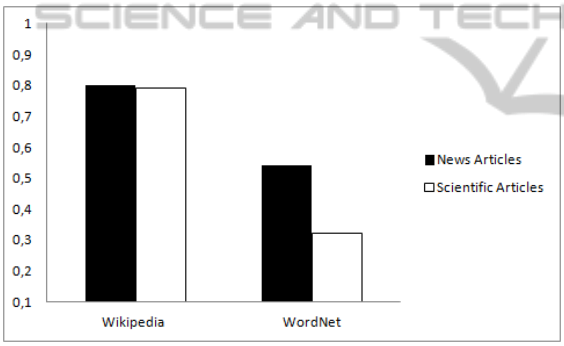

Figure 7: Terms coverage.

5.1.2 Articles Coverage

The search operation made at the beginning of the process returns a set of articles. The created facets are used to navigate subsets of the articles. Depending on the terms coverage and the minimization step of the built hierarchy, links to some articles can be removed. Articles coverage is a very important measure, because it shows the reliability of the built hierarchy. We define the article coverage to be the number of articles accessed by the facets hierarchy divided by the total number of articles.

As illustrated in Figure 8, the Wikipedia hierarchy has nearly 100% articles coverage but the WordNet hierarchy has significantly lower value. The Scientific articles in the WordNet hierarchy has higher value than the news articles because usually scientific expressions related to the user query are always available in the returned articles, as scientific contents are usually more concise than news articles contents. Nevertheless, the Wikipedia hierarchy achieves better results in the articles coverage.

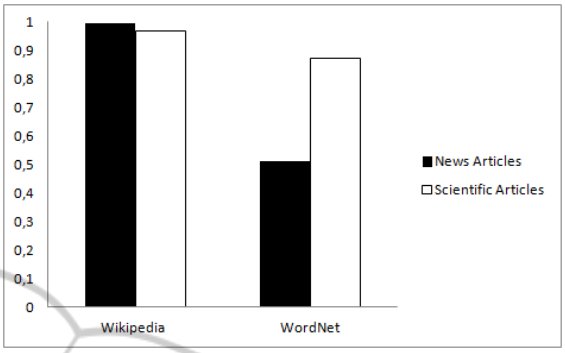

Figure 8: Articles coverage.

5.1.3 Accuracy of Facets

In this assessment, each facet is inspected manually. A value of 0 is assigned to the facet if the facet is not related to the assigned articles, and a value of 1 is assigned to the facet that accurately describes its assigned articles. This measure indicates whether the facets terms extracted from the articles accurately describe the text articles or not. We define the facet accuracy to be the number of facets closely related to the text topic divided by the number of facets assigned to text topics. As shown in Figure 9, the hierarchies have good 87-97% accurate facet terms for both methods and for both data sets.

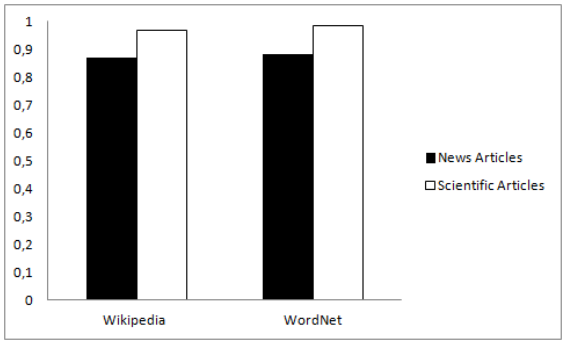

Figure 9: Accuracy of facets.

5.1.4 Significance of the Facet Tree

This assessment is also performed manually. We check whether a facet term is correctly placed in the hierarchy or not. A facet is in a correct place if the parent facet accurately describes its children, and a child facet is placed under the correct parent.

We define the Significance factor to be the number of significant facets in the hierarchy divided by

the number of facets in the hierarchy. Again, all hierarchies have 86-96% facet significance.

We are also interested in the amount of information lost during the translation from Arabic to English and vice versa in the WordNet approach. Again, manually, each term in the hierarchy is assigned a 0 or 1 value for the translation factor. The 0 value is assigned to indicate a bad translation that affected the meaning of the term. The 1 value is assigned to indicate a correct translation. We define the translation factor to be the number of facet terms with exact translation divided by the number of facets in the hierarchy. The translation factor is about 95%; which means a very low level of loss of meanings within the facets terms.

5.2 Comparison of the Facet Hierarchies

In this section, we compare the shapes of the facet hierarchies. A hierarchical facet structure is said to be good if it has a *reasonable number of siblings* for each node and the *expanded path from root to leaf fits in one screen* without scrolling down.

5.2.1 Number of Facets

As shown in Figure 10, the Wikipedia hierarchy has an average of 27 facets for the BBC news articles, while WordNet has only 9 facets. For the Arab Sciencepedia articles, both the Wikipedia hierarchies and the WordNet hierarchies have an average of 15 facets. In the Wikipedia hierarchies, the number of facets decreases from 27 in the news articles to 15 in scientific articles because most of the extracted terms are related to a small number of topics with more concentrated categories as compared to those extracted from the BBC news articles; which *is typical to the nature of related scientific articles as compared to the broad natured news articles*. The WordNet numbers of facets increased, from an average of 9 facets in the BBC news articles to 15 in Arab Sciencepedia articles, because WordNet understands most of the scientific expressions, so better results are obtained.

5.2.2 Average Tree Depth

The average tree depth is computed as follow.

$$average\_depth = \frac{sum\ (\max.\ depth\ of\ each\ facet)}{number\ of\ facets\ in\ the\ final\ facets\ list}$$

The average tree is computed three times for each hierarchy: once for the LingPipe facets, once for the non-LingPipe facets, and once for the whole hierarchy facets.

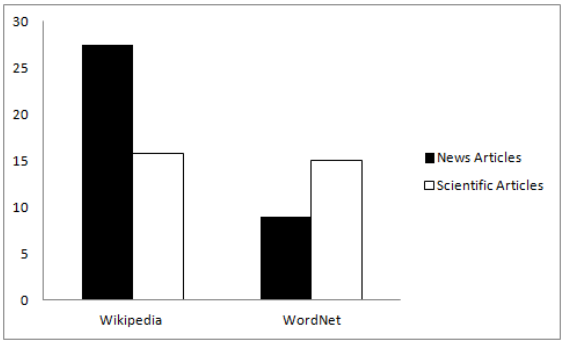

Figure 10: Average number of facets.

The Wikipedia hierarchy predefined facets are flat facets, while the WordNet predefined facets are hierarchical, which makes the WordNet hierarchy depth average better than the Wikipedia hierarchy depth average. The depth average is illustrated by the graph in Figure 11. This is one of the rare cases where the WordNet yields better results than Wikipedia.

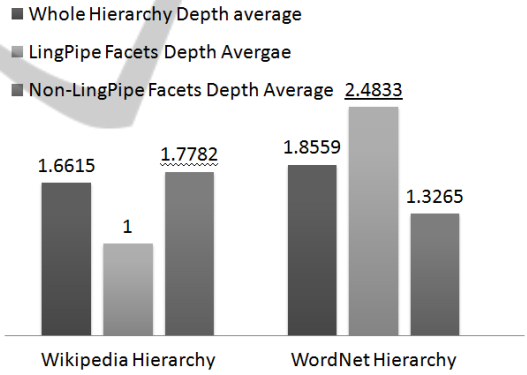

Figure 11: Average depth of facet trees.

5.3 Creation Time

The total time needed for building each of the hierarchies is illustrated in Figure 12. The system configured only to use WordNet hierarchy is always slower than the Wikipedia hierarchy. In the hybrid hierarchy, both hierarchies are created in separate threads. Then a fast merging step is done. So, it is clear that the creation time for the hybrid model is slightly higher than the maximum of each measurement.

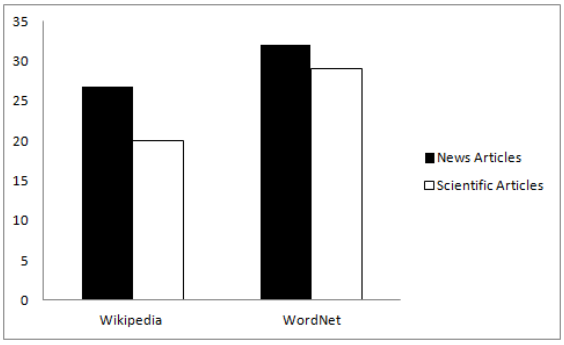

Figure 12: Creation time in seconds.

6 CONCLUSIONS AND FUTURE WORK

In our work, we develop a hierarchical faceted search system for Arabic documents. Building the facets involves two steps: creating the set of facets then building the facets hierarchy. The set of facets is created by extracting the most significant terms from the searched documents. Then two different paradigms are used in the hierarchy building process in order to check whether it is better to use Arabic tools and content or to use English. English tools, such as WordNet, already showed success for creating faceted interface for English content. Using English tools imply a translation step from Arabic to English and vice versa.

The hierarchy built using the Wikipedia categorization is built in a button-up manner. Each term in the candidate terms list is checked against the Wikipedia categorization content, and if found all its parents are inserted into the hierarchy. This step is recursively done for each parent until a three-level hierarchy is obtained.

To use the WordNet IS-A hypernym structure, candidate terms are translated into English. Then, each term is checked against the WordNet structure. Once found, all its parent hierarchy is inserted in the hierarchy under construction. If a term is not found in the WordNet structure but has a tag value equal to one of the predefined facets, it is directly placed under this facet. Any ambiguous term is resolved using the WordNet Domains tool. Finally the terms inserted in the hierarchy are translated to Arabic.

One major drawback appeared while using the English WordNet with Arabic text. Arabic cultural expressions, even if well translated into English, are not covered in the English WordNet content. This results in dropping significant number of candidate terms. Building the facets hierarchy with only a subset of the candidate terms results in low terms and articles coverage. Whereas building the facets hierarchy using the Arabic Wikipedia categorization structure shows better performance in both news articles and scientific articles.

One note must be considered for the usage of the faceted search interface framework with different documents contents. If LingPipe statistical extraction model is used, a training corpus related to the documents contents must be used in order to get acceptable results. LingPipe statistical model tags can be checked versus the English WordNet corpus. Once found, hybrid facets hierarchy is created by merging the two hierarchies giving an enhanced facet hierarchy.

Finally, we are aware that the comparison is the result of a clear competition between a wiki-based approach – and hence organically evolving – with a huge community of contributors versus a well structured approach driven by linguistic science. This benchmarking should be periodically revisited as results would change with the development of better linguistic tools in Arabic language. An interesting extension would be using Arabic WordNet within a few years in the same way it was used to improve QA for Arabic (Abouenour et al., 2008).

REFERENCES

Abouenour L., Bouzoubaa K., Rosso P., 2008. Improving Q/A Using Arabic Wordnet. *In the International Arab Conference on Information Technology (ACIT'2008)*, Tunisia.

Arabic Wikipedia, n.d. [Online] Available at: <http://ar.wikipedia.org/> [Accessed 13 8 2012].

Arabic Wikipedia Categorization, n.d. The Arabic Wikipedia Categorization root category, [Online] Available at: <http://ar.wikipedia.org/wiki/تصنيف:محتويات_ويكيبيديا> [Accessed 20 6 2012].

Arabic WordNet, n.d. [Online] Available at: <http://www.globalwordnet.org/AWN/> [Accessed 7 6 2012].

ArabSciencepedia, n.d. [Online] Available at: http://www.arabsciencepedia.org/ [Accessed 7 7 2012].

BBC Arabic, n.d. [Online] Available at: <http://www.bbc.co.uk/arabic/> [Accessed 7 7 2012].

BinAjiba, Y., n.d. *Arabic NER Corpus and Documents.* [Online] Available at: <http://www1.ccls.columbia.edu/~ybenajiba/downloads.html> [Accessed 12 2 2013].

Bing Translator, n.d. *The Bing translator API* [Online] Available at: <http://www.microsoft.com/web/post/using-the-free-bing-translation-apis> [Accessed 4 9 2012].

Dakka, W. & Ipeirotis, P. G., 2008. Automatic extraction of useful facet hierarchies from text databases. In *IEEE 24th International Conference on Data Engineering (ICDE), pp.466-475.*

Google API, n.d. *Google Custom Search API* [Online] Available at: <http://code.google.com/ apis/ customsearch/> [Accessed 23 2 2013].

GSON, n.d. *Google GSON API* [Online] Available at: <https://code.google.com/p/google-gson/> [Accessed 13 2 2013].

Hearst, M. A., 2008. UIs for faceted navigation recent advances and remaining open problems. *In Workshop on computer interaction and Information retrieval, HCIR,, Redmond, WA.*

Hearst, M. A. & Pedersen, J. O., 1996. Reexamining the cluster hypothesis.In *Proceedings of the 19th Annual International ACM/SIGIR Conference, Zurich.*

JSON, n.d. *The Json file format* [Online] Available at: <http://www.json.org/> [Accessed 14 3 2013].

JSOUP, n.d. *Java HTML parser API* [Online] Available at: <http://jsoup.org/> [Accessed 23 2 2013].

Lim, D. et al., 2011. Utilizing wikipedia as a knowledge source in categorizing topic related korean blogs into facets. In *Japanese Society for Artificial Intelligence (JSAI), Takamatsu.*

LingPipe, n.d. LingPipe Named Entity Recognizer, [Online] Available at: http://alias-i.com/lingpipe/ demos/tutorial/ne/read-me.html [Accessed 4 9 2012].

Manning, C. D., Ranghavan, P. & Schütze, H., 2009. *Introduction to Information Retrieval.* Cambridge University Press.

Mark, S. & Croft, B., 1999. Deriving concept hierarchy from text. *In proceeding of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*.

Naver, n.d. *Naver Open API* [Online] Available at: <http:// dev.naver.com/openapi/> [Accessed 15 2 2013].

Wikipedia Redirect, n.d. [Online] Available at: <http://en.wikipedia.org/wiki/Help:Redirect> [Accessed 15 2 2013].

Stoica, E. & Hearst, M. A., 2006. Demonstration :Using WordNet to build hierarchical networks. *In the ACM SIGIR Workshop on Faceted Search.*

Stoica, E. & Hearst, M., 2004. Nearly-Automated Metadata hierarchy creation. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Boston.

Stoica, E., Hearst, M. & Richardson, M., 2007. Automating creation of hierarchical Faceted metadata structures. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Rochester NY, USA.

Tripathi, S. & Sarkhel, J. K., 2010. Approach to machine translation. *In The Americas Lodging Investment Summit (ALIS) Vol.57, pages 388-393*.

Tunkelang, D., 2009. *Faceted Search.* Morgan & Claypool.

Wikipedia Categorization, n.d. [Online] Available at: <http://en.wikipedia.org/wiki/Wikipedia:Categorization> [Accessed 12 5 2013].

WordNet Domains, n.d. [Online] Available at: <http://wndomains.fbk.eu/> [Accessed 13 2 2013].

WordNet, n.d. [Online] Available at: <http:// wordnet.princeton.edu/> [Accessed 3 2 2013].

Yahoo History, n.d. [Online] Available at: <http:// docs.yahoo.com/info/misc/history.html> [Accessed 9 8 2012].

Yan, N. et al., 2010. *FacetedPedia: dynamic generation of query-dependent faceted interface for wikipedia. In Proceedings of the 19th international conference on World wide web, pp.651-660.*