**Text Clustering and Categorization**

**Dr. Racha Elkashef**

**Eng. Ehab Abdelhamid**

**Eng. Michael Azmy**

**Dr. Aly Fahmy**

**Dr. Ahmed Rafea**

**February 15, 2010**

**Technology: Clustering and Categorization**

1. **Brief Overview**

Analysis of data can reveal interesting, and sometimes important, structures or trends in the data that reflect a natural phenomenon. Discovering regularities in data can be used to gain insight, interpret certain phenomena, and ultimately make appropriate decisions in various situations. Finding such inherent but invisible regularities in data is the main subject of research in data mining, machine learning, and pattern recognition.

Data clustering is a data mining technique that enables the abstraction of large amounts of data by forming meaningful groups or categories of objects, formally known as clusters, such that objects in the same cluster are similar to each other, and those in different clusters are dissimilar. A cluster of objects indicates a level of similarity between objects such that we can consider them to be in the same category, this simplifying our reasoning about them considerably.

Text categorization is defined as the process of assigning predefined class labels to new text documents based on what the classifier learnt from the training set of documents.

1. **State of the Art (For Latin Languages)**
   * Technology and Future Trends

Clustering approaches can be classified along different independent dimensions. For instance, different starting points, methodologies, algorithmic point of view, clustering criteria, and output representations, usually lead to different taxonomies of clustering algorithms. Different properties of clustering algorithms can be described as follows:

1. *Non-hierarchical methods:*

* k-means and its extensions (spherical kmeans, kernel kmeans and bisecting k-means)
* Buckshot
* The leader-follower algorithm
* Self-organizing map (SOM)

1. *Hierarchical methods:*

* Agglomerative
* Divisive

1. *Generative algorithms:*

* Guassian model
* Expectation Maximization
* Von Mises-Fisher
* Model-based k-means

1. *Spectral Clustering:*

* Divide & merge
* Fuzzy coclustering

1. Density-Based Clustering: The rationale of density-based clustering is that a cluster is composed of well-connected dense regions. DBSCAN is a typical density-based clustering algorithm, which works by expanding clusters to their dense neighborhood (Ester, Kriegel, Sander, & Xu, 1996).
2. *Phrase based models:*
   * + - Suffix tree clustering
       - Document index graph

Major categorization techniques are K nearest neighbor (KNN), support vector machines (SVM), Naïve Bayesian (NB), Naïve Bayesian Multinomial (MNB), Decision Tree, hidden Markov model (HMM), maximum entropy, and neural network (NN).

The following modules are generally used in any clustering technique:

* Preprocessing and Feature extraction
* Dimensionality reduction (e.g. Principal component analysis, Nonnegative matrix factorization, Soft spectral coclustering, Lingo)
* Similarity Measure
* Clustering algorithm
* Evaluation using Internal and external validity measures

To perform the text categorization task, a set of modules have to be applied. Table 1, shows the most famous text categorization algorithms alongside the different needed modules.

Table 1: Text categorization components

|  |  |  |  |
| --- | --- | --- | --- |
| Modules | **SVM** | **KNN** | **MNB** |
| Feature Extraction  (Stemming & stop word removal) | Yes | yes | yes |
| Feature Selection | Yes | yes | yes |
| Document Representation | Yes | yes | yes |
| Learning | Yes | yes | yes |

* + Applications and Reported Performance

Clustering is used in a wide range of applications, such as marketing, biology, psychology, astronomy, image processing, and text mining.

Document clustering techniques are widely used in:

* Information retrieval
  + Improve precision and recall
  + Organizing results
  + Example online search engines: clusty.com and iboogie.com
* Organizing documents
  + Finding nearest documents to a specific document
  + Automatically generate hierarchical clusters of documents
  + Browsing a collection of documents, corpus exploration
* Indexing and Linking

Text categorization techniques are used in many applications, including:

* E-mail filtering,
* Mail routing
* Spam filtering
* News monitoring
* Sorting through digitized paper archives
* Automated indexing of scientific articles
* Classification of news stories
* Searching for interesting information on the WWW
* Classify business names by industry
* Classify movie reviews as Favorable, Unfavorable and Neutral
* Classify web sites of companies by Standard Industrial Classification (SIC) code.
  + Available packages for clustering and Categorization
* CLUTO: software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. Clustan: Clustering and clustering analysis software. Matlab Clustering Package.
* Weka workbench: is an open source Data Mining software package written in Java and it is freely available : [www.cs.waikato.ac.nz/~ml/weka/index.html](http://www.cs.waikato.ac.nz/~ml/weka/index.html).
* Minorthird: toolkit for extraction and classification of text: <http://minorthird.sourceforge.net>

1. **State of the Art (For Arabic Language)**
   * Technology and Future Trends

Yet, No Arabic document clustering technique has been developed for Arabic, some feature reduction and stemming techniques are only found.

* + Current and Envisioned Applications and Market Priorities
    1. Arabic Document Clustering in E-Learning
    2. Arabic Clustering in Press and Media Analysis
    3. Arabic Clustering and Categorization in Social Networks Analysis

1. **Dependency Between Technologies**
   * Clustering and Feature Selection (e.g. use of frequent itemsets and closed frequent itemsets)
   * Clustering and Social Networks Analysis
   * Clustering and Image Processing
   * Clustering and Outliers Detection
   * Clustering and Gene Expression Analysis, etc
   * Clustering and semantic based techniques

Table 2 shows a comparison between the different clustering techniques.

Table 2: Comparison between Clustering Techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Technique | **Dis(Similaity)**  **Matrix** | **Number of**  **Clustering** | **Sensitive to**  **Outliers** | **Shape of**  **Clusters** |
| *Non-hierarchical methods:* | NO | Known | Yes | Spherical-Shaped Clusters |
| *Hierarchical methods* | Yes | Known | Not as Non-Hierarchical approaches | Elongated –Shaped Clusters |
| *Generative algorithms* | No | Known | Yes | Spherical-Shaped Clusters |
| *Spectral Clustering* | No | Known | No | Arbitrary-Shaped |
| Density-Based Clustering | Yes | Unknown | No | Arbitrary shaped-Clusters |
| Phrase-Based Models | Yes | Known | No | Arbitrary shaped-Clusters |

1. **Arabic Language: An Overview**

Arabic language is a highly inflected and a non-concatenative language; it has much richer morphology than English. The majority of words have a tri-letter root. The rest have either a quad letter root, penta-letter root or hexa-letter root.

The ignorance of these morphological features and the special characteristics of the Arabic language make the existing text categorization techniques more sensitive to noise when used to categorize Arabic texts which makes the categorization of these texts is a challenging task. The language dependent components are those components that are related to the extraction of the morphological features and the special characteristics of the Arabic language. Examples of these components are stemming, POS tagging, and stop word removal. Although there are three approaches for stemming for the Arabic language (which are, root-based stemmer, light stemmer, and statistical stemmer), the light stemmer superseded the two other approaches in some of the recent research works.

1. **Language Resources (datasets, benchmarks)**
   * Available Resources (English, Arabic)
     + English Corpora: Reuters-21578, Yahoo News, ACM Magazines, UW Dataset, WAP from WebAce, WebKB, 20-Newsgroups corpus, Cade corpus, Brown corpus.
     + Arabic Corpora:
       - free non-standard web documents
       - In-house collected corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, Al-Dostor, El Akhbar, El Gomhoria, and the Arabic version of the Food and Agriculture Organization (FAO) website as a source for the agriculture category articles.
   * Needed Resources (English, Arabic)
     + TREC Arabic dataset : TREC-5, TREC-6, and TREC-7
     + W0030 : Arabic Data Set
     + Levantine Arabic QT Training Data
     + The Arabic NEWSWIRE corpus of LDC

Table 3, shows the most famous text categorization algorithms alongside the different needed resources.

Table 3: Text categorization resources

|  |  |  |  |
| --- | --- | --- | --- |
|  | **SVM** | **KNN** | **MNB** |
| Category annotated documents | Yes | Yes | yes |
| Stop word list | Yes | Yes | yes |
| Morphology analyzer | Optional | Optional | Optional |

1. **Strengths, weaknesses, opportunities and threats** 
   * Strengths
     1. Arabic is a native language
     2. Some free web documents
     3. Yet, No Arabic document clustering technique has been developed for Arabic, some feature reduction and stemming techniques are only found.
   * Weaknesses
2. There are no free standard datasets
3. No available benchmarks
   * Opportunities
     1. The market is in a bad need for that application
   * Threats
     1. Some companies and research Labs, like Sakher, IDI, Google, DCU Research Lab, and Texas University Research Lab have already done some product for Arabic categorization, with acceptable performance.
4. **Suggestions for Survey Questionnaire**
5. **List of people/organizations pioneers in each application area to be targeted by the Survey**

### National: People that have a research work in the area of Arabic text categorization:

1. Sakhr company, created its Arabic text categorizer in 2004:

<http://www.sakhr.com/Technology/Categorization/Default.aspx?sec=Technology&item=Categorization>

1. Hisham El-Shishiny, created his Patent in Arabic text categorization in 2005:

"Method and system for categorizing Arabic text", <http://www.wipo.int/pctdb/en/wo.jsp?wo=2005015434>

* Arab: People that have a research work in the area of Arabic text categorization:

1. Rehab Duwairi:

Department of Computer Information Systems, Jordan University of Science and Technology, Irbid, Jordan.

1. Riyad Al-Shalabi:

Amman Al-Ahliya University, Jordan.

1. Abdelwadood. Moh’d. Mesleh:

Computer Engineering Department, Faculty of Engineering Technology, Balqa’ Applied University, Amman, Jordan.

1. Mohamed El Kourdi:

School of Science & Engineering Alakhawayn University

1. Hassan Sawaf:

computer science department VI RWTH Aachen Ahornstrasse

### International: The key names of the international people that have a research work in the area of text categorization:

1. Fabrizio Sebastiani
2. David D. Lewis
3. Thomas Hofmann
4. Yiming Yang
5. Fabio Crestani
6. Soumen Chakrabarti
7. William W. Cohen
8. Stefan Wermter
9. Yiming Yang
10. Shuigeng Zhou
11. Mohamed Kamel
12. **Key persons in each application area (on technical/LR levels)** 
    * **The same as in section 9**
13. **Suggestions for Language Resources (specific to the application area) if ALTEC would like to start collection immediately.** 
    * In-house collected corpus from online Arabic newspaper archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, Al-Dostor, El Akhbar, El Gomhoria, and the Arabic version of the Food and Agriculture Organization (FAO) website as a source for the agriculture category articles.
14. **Summary**