ADD: Arabic Duplicate Detector
A Duplicate Detection Data Cleansing Tool

Ramzi A. Haraty, Ralph Varjabedian
Computer Science Program, Lebanese American University
P.O. Box 13-5053 Chouran
Beirut, Lebanon 1102 2801
Email: rharaty@lau.edu.lb

Data mining is a relatively new term; it was introduced in the 1990s. Data mining is the process of extracting useful information from huge amounts of data. It is sometimes referred to as "data discovery" or "knowledge discovery" in databases. What exactly defines useful information depends on the goal that data mining was for in the first place. Useful information can be used to increase revenue and to cut costs. It can also be used for the purpose of research. Advances in hardware and software in the late 1990s made data centralizing possible. Data centralizing is also called "data warehousing" or "data warehouse for the centralized data". With the process of data centralization came a very important issue, the quality of the data that has been centralized, since centralization includes the joining of multiple data sources. The data given as an input for the data mining process should be of high quality in order for the results of the data mining process to be accurate and reliable. Before data could be mined to extract useful information, it goes through a process called data cleansing. This process is as old as the word "data" itself; however, the term regained significance in the 1990s. Data cleansing involves several steps and processes that include one or more algorithms. This paper addresses one important step, which is duplicate data detection. The paper presents a duplicate detection method called the Efficient *k*-way Sorting Method. The paper also presents a tool called Arabic Duplicate Detection, which is based on our method and is tailored for Arabic data.

Keywords
Arabic data, data cleansing, duplicate detection, and knowledge based-systems.