Chapter I
Introduction

The Word Wide Web is the largest known collection of hypertext data, it is dynamic in nature and grows exponentially with time; for example, the number of documents published on the web reached 550 billion documents in 2001 [1]. This vast quantity of data makes finding proper information a tedious and time consuming activity for web users.

Most web users are using search engines to find their desired information on the web, however, search engines usually retrieve large number of web pages as a result of a user query; most of these pages are irrelevant to the user. Typically users browse through the first two pages of the returned list of pages. Search engines employ ranking algorithms that aim to placing the most relevant pages to the user query into the top of the list of returned pages.

Many recommender systems (i.e. systems that aim to enhance search results to satisfy user's queries) which are based on user feedback were developed to solve the aforementioned problem, but all these techniques use explicit user feedback which is costly in terms of time and resources.

An implicit user feedback can be driven by using different features like click-through, time spent on the page (browsing time), exit type, added to favorites, and scrolling count [2]. Clustering can be employed by search engines to group together similar pages; subsequently these clusters can be utilized in ranking pages or in importing the presentation styles of search engines.

1.1 Knowledge Discovery and Data Mining:

Data mining is the process of sophisticated extraction of novel, valid patterns and relationships in large databases [3]. Data mining can be applied to different data representations such as numeric, textual, and multimedia. Currently there are wide ranges of data mining applications which include association, sequence or path analysis, classification, clustering, and forecasting.

Data mining is just one step of a large process known as Knowledge Discovery in Database (KDD), as shown in Figure 1.1. KDD has several steps namely: *data cleaning* to remove noise and inconsistent data, *data integration* to integrate data when multiple data sources may be combined, *data selection* to extract relevant data to the analysis task, *data transformation* to transform data into forms appropriate for mining, *data mining* by adding intelligent methods to extract data patterns, *pattern evaluation* to test interesting patterns representing knowledge, and *knowledge presentation* to present mined knowledge to the user in clear and understandable way. This thesis is concerned with a special type of data mining application which is web clustering application.

1.2 Motivation

As a result of the continuous increase in the amount of available information on the web, users can't find their needed information easily. As of December 1998, 85% of web users located their desired information by using search services and 60% of them used web directories [4]. So, web search is the most important service for web users, but there is a big problem in search engines; the problem is that when a user searches for a topic, thousands of web pages are returned as a result to the user's query but few of them are relevant to the user's query.

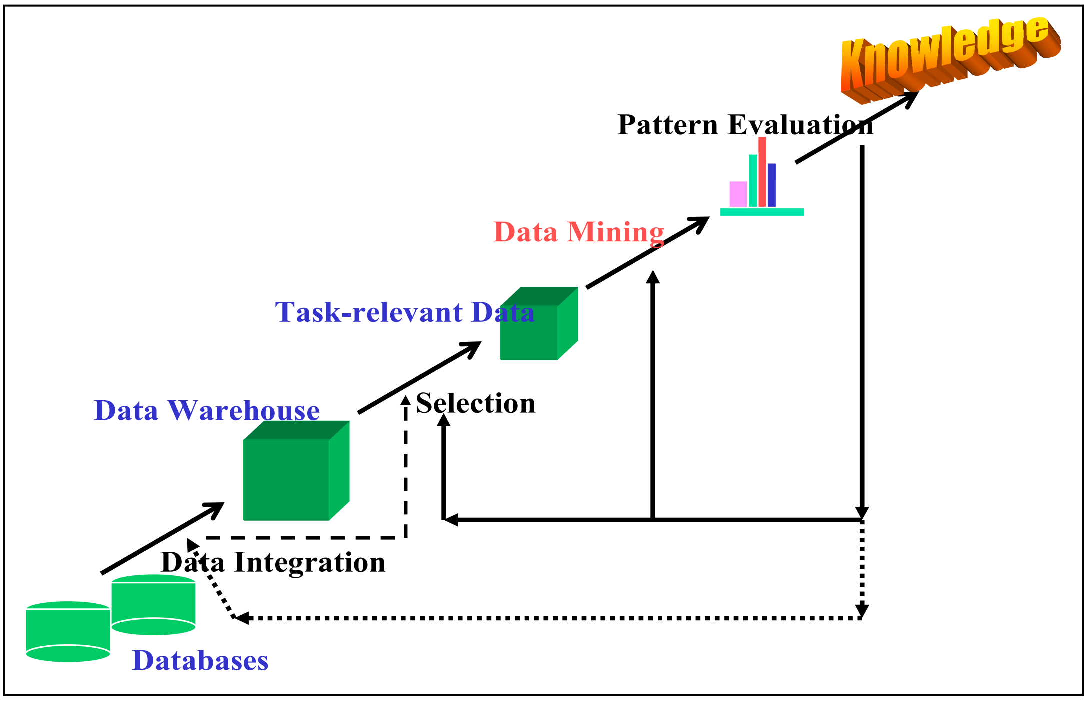
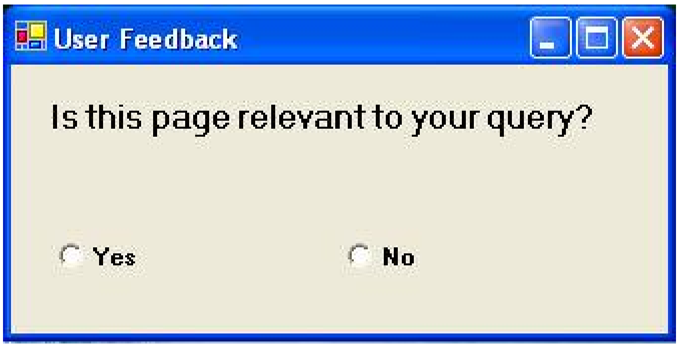


Figure 1.1. The phases of the KDD process.

Many techniques and recommender systems were developed to enhance user search precision; most of these techniques depend on user feedback. User feedback is divided into two main types:

- *Implicit user feedback*: where user feedback is driven from user's behaviors like browsing time, scrolling count, exit type, and added to favorites; furthermore as described in [2] these implicit measures can be used as alternatives to explicit user feedback.
- *Explicit user feedback*: where user feedback is taken by asking the user to fill in a feedback form like the one shown in Figure 1.2. This type of user feedback is time consuming and annoying to users.



Is this page relevant to your query?

○ Yes ○ No

Figure 1.2. An example of an explicit user feedback form

1.3 Thesis Methodology

In this thesis we investigate user browsing time as an implicit user feedback measure to enhance web search accuracy by enhancing the quality of web pages' clusters. The objective is to enhance web search precision based on user browsing time. The experiment shows an incentive enhancement on search precision after applying the proposed technique.

1.4 Thesis Outline

This thesis is organized into 5 chapters. The first chapter has introduced the objectives and motivations of this research. Chapter 2, by comparison, introduces recent research related to this work; this chapter consists of two main sections; the first one describes clustering techniques, web clustering approaches, and document clustering algorithms. The second section, by contrast, compares the different forms of user feedback widely used in research and industry.

Chapter 3 presents the proposed technique. This chapter is divided into five main sections. The first section describes the dataset properties (the one that is used for experiments), the second section, explores the clustering algorithm that is used to create the initial clusters in the proposed framework, the third section explains in details the filtering phase of this technique. The fourth section demonstrates the re-clustering phase;

finally, the last section reviews the measures of cluster quality that are used to evaluate the proposed technique's results.

Chapter 4 presents the experimental results of the prototype that are developed to assess the benefits of employing user browsing time as a mechanism of enhancing cluster quality. This is done by evaluating clusters' quality before and after taking user browsing time into consideration. Chapter 5 summarizes the contributions of this thesis and highlights future work.