DEDICATIONS

إلى من علمي شموخ الشمس قبل فراعة الفقيد...
إلى الشعمة التي تعترف بئنبر لي جريدة الحياة الشريفة...
إلى والدي العزيز

إلى الهيئ uniq السماه في غل مساء
تحوى لي بالحفظ والتوافق من الله...
إلى من أحبني أهمن عينيما إتجلالا وإشباهاً
إلى أمي العبيبة

إلى القيقب الذي إحتوائي في أقصى ضوءه نادي...
إلى من علمني معرفته أتعلاني على نبياني...
إلى أخي و فخذي... رحمر

إلى البسند الواحد الذي يامل لألامي و يعمر للحري...
إلى إحوائي وأحويائي غل في مطامه

إلى التي تعرسم في بجود الإيمان
ورغمها حتى آتيت أطلالما
إلى مهلة الإتحاد الإسلامية

إلى أمّ "إفرا".
التي مئات الأرض نوراً وعلماً
إلى أمتي الفاخرة... أمة الإسلام

أهدي هذا العمل سالماً لله، أن يجعله علماً نافعاً.
ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Rehab Duwairi for her support, sharing of knowledge and the opportunities that she gave me. She has been the most generous with her time to read this thesis carefully and make insightful comments and suggestions.

I would like to thank especially Mr. Benjamin Fung who helped me in understanding the source code of the FIHC clustering algorithm. Many people have helped and contributed their time to this research of this thesis. Many thanks go to Mr. Hassan Alzoubi, Mr. Mohamad Fayez, and all students who participated in the experiment of our technique.

TABLE OF CONTENTS

ii
I Introduction ..................................................................................................... 1
  1.1. Knowledge Discovery and Data Mining ............................................. 2
  1.2. Motivation ......................................................................................... 2
  1.3. Thesis Methodology ................................................................. 4
  1.4. Thesis Outline ............................................................................. 4
II Background and Related Work .............................................................. 6
  2.1. Clustering .................................................................................... 6
    2.1.1. Clustering Techniques ......................................................... 6
    2.1.2. Web Clustering Approaches ............................................... 8
    2.1.3. Document Clustering Algorithms ...................................... 9
    2.1.4 Clustering Applications ....................................................... 16
  2.2 User Feedback ............................................................................ 17
    2.2.1 Relevance Feedback in Information Retrieval ....................... 17
    2.2.2 Implicit User Feedback ......................................................... 20
    2.2.3 Semi-Supervised Clustering ................................................. 20
    2.2.4 Feedback-Driven Document Clustering ................................. 22
III A User-Enhanced Clustering Technique ............................................. 24
  3.1 Creating Initial Clusters .............................................................. 24
  3.2 Filtering ....................................................................................... 25
  3.3 Re-Clustering ............................................................................. 26
3.4 Assessment of Cluster Quality........................................... 27
3.5 Dataset Characteristics..................................................... 28

IV Experimentation and Result analysis..................................... 30

4.1 The Data Set........................................................................ 30
4.2 Experiments........................................................................ 32
4.3 Result Analysis.................................................................... 36

V Conclusions and Future Works.............................................. 43

References............................................................................... 45

Arabic Abstract......................................................................... 48

**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>iv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>The phases of the KDD process</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>An example of an explicit user feedback form</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>A pictorial representation of hierarchical clustering</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>An example of $q_{\text{keyword}}$, $q_{\text{in}}$, and $q_{\text{out}}$ vectors</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>The FIHC Algorithm</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>An illustration of semi-supervised clustering</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>A screen shot of the Genius Filter</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>An Example of the first two phases in the proposed technique</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Relationships between programs used to create the data set</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>An MS SQL server view of tables in the database</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Average precision at each candidate threshold value</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>The F-measure values of the initial clusters</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Precision of the clusters before and after filtering</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>F-measure values before and after filtering</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>Enhancements of clusters’ quality after re-clustering with $K=3$ and $N=5$</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>Enhancements of clusters’ quality after re-clustering with $K=5$ and $N=8$</td>
<td></td>
</tr>
<tr>
<td>4.9</td>
<td>Enhancements of clusters’ quality after re-clustering with $K=9$ and $N=10$</td>
<td></td>
</tr>
<tr>
<td>4.10</td>
<td>A Comparison between the average of F-measure values with different $K$ and $N$ values</td>
<td></td>
</tr>
<tr>
<td>4.11</td>
<td>A comparison of F-measure values at $K=3$ and $N=5$ with and without optimization</td>
<td></td>
</tr>
<tr>
<td>4.12</td>
<td>A comparison of F-measure values at $K=5$ and $N=8$ with and without optimization</td>
<td></td>
</tr>
<tr>
<td>4.13</td>
<td>A comparison of F-measure values at $K=9$ and $N=10$ with and without optimization</td>
<td></td>
</tr>
</tbody>
</table>
4.14 A comparison between the average of F-measure values with different $K$ and $N$ values 42
4.15 Final results of cluster's quality enhancement 42

LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.14</td>
<td>A comparison between the average of F-measure values with different $K$ and $N$ values</td>
<td>42</td>
</tr>
<tr>
<td>4.15</td>
<td>Final results of cluster's quality enhancement</td>
<td>42</td>
</tr>
</tbody>
</table>
2.1 Implicit measures of user feedback

4.1 Symbols that appear in URLs and their corresponding symbols in Windows

4.2 Distribution of pages to classes

4.3 Distribution of documents to clusters after applying the FIHC algorithm

4.4 Database tables, their fields and purposes
ABSTRACT
ENHANCING CLUSTER QUALITY BY USING USER BROWSING TIME

By
Khaleifah S. Al.jada'
Chairperson: Dr. Rehab Duwairi

The World Wide Web currently contains billions of documents; this causes difficulty in finding the desired information by users. Many search engines come out to help users finding their desired information but all search engines still return hundreds of irrelevant web pages that do not fulfill the user's query.

Many search engines use clustering to group documents that are relevant to the user's query before returning them to the user, but there is no document clustering algorithm that has an accuracy that can prevent retrieving irrelevant documents. Relevance feedback is a classic information retrieval technique, it is employed to enhance clustering results by using two different user feedback forms namely, explicit user feedback and implicit user feedback. Currently most available systems use explicit user feedback but most recent research rely on the substitution of explicit user feedback with implicit user feedback.

In this thesis, the researcher has introduced a new technique to enhance cluster quality by using user browsing time as an implicit measure of user feedback, rather than using explicit user feedback as in all previous research and techniques. The major contributions of this work are: investigating user browsing time as an implicit measure of user feedback and proving its efficiency, enhancing cluster quality by using a new clustering technique that is based on user browsing time, and developing a system that tests the validity of the proposed technique.

Experimental and performance studies have demonstrated the effectiveness and efficiency of the proposed technique in comparison to one of the best document clustering algorithm currently used.