Contents

1

Thesis Committee 2

Abstract 3

Acknowledgements 4

Publications related to this thesis 5

1 Introduction 1

1.1 Motivation 1

1.2 The query log mining process 6

1.3 Contributions 10

1.4 Thesis Outline 12

2 State of the Art 14

2.1 User's click data analysis 16

2.2 Query association methods 19

2.3 Query clustering methods 20

2.4 Query classification methods 24

2.5 Conclusions 25

3 User's Click Data Analysis 27

3.1 User's click data pre-processing 27

3.2 DataSets 29

3.3 Sessions and Queries 31

3.4 Keyword Analysis 31
3.5 Click Analysis 34
3.6 Query Session Models 36
3.6.1 Aggregated query session model 36
3.6.2 Markovian query session model 40
3.6.3 Time distribution query session model 44
3.7 Conclusion 47

4 Query Association Methods 49
4.1 The query term based method 49
4.1.1 The method 49
4.1.2 Experimental results 51
4.2 The co-citation method 54
4.2.1 The method 54
4.2.2 Experimental Results 56
4.3 Conclusion 61

5 Query Clustering Methods 64
5.1 The method based on document terms 64
5.1.1 Application: Query Recommendations 66
5.1.2 Application: Document Recommendations 67
5.1.3 Experimental results 68
5.2 The method based on snippets 74
5.2.1 The bias reduction method 76
5.2.2 Experimental Results 78
5.2.3 Clustering Process 78
5.2.4 Distance functions comparison 80
5.2.5 Evaluation of recommendation algorithms 84
5.2.6 Query Recommendation 86
5.2.7 Answer Ranking 86
5.3 Conclusion 89

6 Query Classification Methods 92
6.1 The classification method 92

6.1.1 Preliminaries 92
6.1.2 The classification method 94
6.1.3 Experimental results 96
6.1.4 Conclusion 100
6.2 The Web directory maintenance method 101
6.2.1 The method 101
6.2.2 Experimental Results 104
6.3 Conclusion 111

7 Conclusions 112
7.1 Concluding Remarks 112
7.2 Future work 117

Bibliography 119

List of Tables

2.1 Summary about Related Work in Web Usage Mining 18

2.2 Summary about Related Work in Query Clustering 23

3.1 Characteristics of L_1, L_3 and L_6. 30

3.2 Statistics that summarize the contents of L_3. 30

3.3 (A) The top 10 queries. (B) List of the top 10 query terms sorted by the number of query occurrences in the log file. Queries were written originally in Spanish. 33

3.4 Expected values (in seconds) for the Weibull distributions involved in the query session model. 46

3.5 Value of the Kolmogorov statistic D_n for each transition. 47

4.1 Examples of "Quasi-synonym" queries recommend each other. 60

4.2 Queries used for the experiments and recommendations with strong levels of consistency, sorted by relevance. 63

5.1 Statistics for the clusters found for $k = 600$ 79

5.2 Distance functions comparison based on a Mantel's test. 81

5.3 Distance distributions for a set of 600 pairs of quasi-synonym queries. Results are shown in percentages. 83

5.4 Medians for each decile of the distance distributions shown in table 5.3. 84

5.5 Distance functions comparison based on quasi-synonym queries sorted by distance. The columns labeled with $P\%$ indicate the percentile in the distance function distribution 85

5.6 Queries selected for the experiments and the cluster to which they belong for the solution obtained using the unbiased snippet terms based method for $k = 600$. Results are sorted by their cluster rank. Proper nouns are showed in italics. 91

6.1 Queries selected for the evaluation of the method of query classification in directories sorted by distance. 98
6.2 Categories of the TodoCl directory selected for the experiments. 105
6.3 Precision and Recall. 107
6.4 Top-5 queries recommended for each category. 110

List of Figures

1.1 The user feedback cycle. 5

1.2 Log file sample in common log format. 6

1.3 The proposed query log mining process 8

1.4 Relations between the data mining techniques used in this thesis and the applications generated 10

3.1 Data model used in the query log database. 28

3.2 The phases of the text pre-processing used in this thesis. 29

3.3 (A) Number of new queries registered in the logs. (B) Log plot of number of queries v/s number of occurrences. 32

3.4 (A) Log-plot of the query term collection. (B) Log-plot of the query terms that do not appear in the text collection. (C) Log-plot of the text terms that do not appear in the query collection. (D) Log-plot of the overall text term collection. 34

3.5 Scatter plot of the intersection between query term and text term collections. 35

3.6 (A) Log plot of number of selections per query session. (B) Log plot of number of selections per position. 36

3.7 Predictive model for the number of clicks in a query session with a known number of queries formulated. 38

3.8 Contingency table and class distribution for the joint events number of queries made and number of pages selected. 39

3.9 Markovian transition query session model. 42

3.10 Time distribution query session model. 46

4.1 Recommendations based on query terms associated to the concept *computation*. . . . 52

4.2 Recommendations based on query terms associated to the concept *law*. 53

4.3 Comparison of the ranking of two queries. 55
4.4 Query `Valparaiso` and associated recommendations. 57
4.5 Query `maps` and associated recommendations 58
4.6 Query `Naval battle in Iquique` and associated recommendations. 59
4.7 Query `fiat` and associated recommendations. 61
4.8 Distribution of the opinions for the both questions. 62

5.1 Cluster quality vs. number of clusters. 69
5.2 Clusters for the experiment, sorted by their cluster rank. 71
5.3 Ranking of queries recommended to the query *Chile rentals*. 72
5.4 Average retrieval precision for the queries considered in the experiments. 72
5.5 Average retrieval precision for the documents considered in the experiments. 73
5.6 Plot of the logarithm for the Rank of last URLs clicked in query sessions versus the logarithm for the number of users. 77
5.7 Quality of clusters vs. number of clusters. 79
5.8 Histogram of the number of queries per cluster for the (A) snippet terms based method and the (B) unbiased snippet terms based method. 80
5.9 Running Time for the Clustering Processes. 80
5.10 (A) Histogram of the distances for the query terms based function. (B) Histogram of the distances for the co-citation based function. (C) Histogram of the distances for the snippet terms based function. (D) Histogram of the distances for the unbiased snippet terms based function. 82
5.11 Average retrieval precision of the query recommendation algorithms. 87
5.12 Average retrieval precision of the proposed ranking algorithm. 88
5.13 Scatter plot of the original ranking versus the proposed ranking for the 300 documents considered in the experiments. 89

6.1 The hierarchical classification schema proposed 96
6.2 User opinions for the experiments based on query taxonomy classification. Figure A) shows the results for the hierarchical classifier and Figure B) shows the results for the flat classifier. 97
6.3 Average precision of the retrieved documents for the methods based on classifiers and for the search engine ranking. 100
6.4 Histogram of distances of query sessions to their assigned categories. 106

6.5 Average precision for the two methods over the top ten results. 107
6.6 Average precision for the query recommendation method over the top ten results. . 108
6.7 Histogram of estimated precision of documents in the manual directory. 109