Unsupervised Identification of the User's Query Intent in Web Search

Liliana Calderón-Benavides

TESI DOCTORAL UPF / ANY 2011

DIRECTOR DE LA TESI
Dr. Ricardo Baeza-Yates
Departament of Information and Communication Technologies

UNIVERSITAT
POMPEU FABRA

To Silvia

Acknowledgments

I would like to thank my advisor, Ricardo Baeza–Yates, for his supervision during this work, for giving me so much freedom to explore and discover new areas of Web Mining, and for his invaluable support in the various stages I have lived in this process. I thank to Mari–Carmen Marcos for the numerous stimulating advices and discussions, they have been of great value.

I thank my friends and colleagues at UPF with whom I have had the pleasure of working over the years. These include David Nettleton, Christian Middleton, and all the members of the WRG and Yahoo! Research Barcelona. I would like to thank Vicente Lopez and Joan Codina from Barcelona Media, I learnt very much from our work together. I thank Devdatt Dubhashi for his invitation to work together at Chalmers University of Technology / Göteborg University, and together with him to Libertad Tansini and Vinay Jethava for their contributions and advices to improve this work. I appreciate immensely to Vassilis Plachouras and Shirish Tatikonda for their comments and reviews in the writing of this thesis. Special thanks to people from the DTIC, in particular to Josep Blat, Xavier Binefa and Judith Champion for their collaboration every time that I needed it.

I would also like to take this opportunity to thank to the UNAB represented by Alberto Montoya for its continued support to realize my dreams.

I am grateful to all my friends in the amazing city of Barcelona, for being the surrogate family during the years I stayed there and for their continued moral support thereafter. I am also grateful to Francesca, Cesar, Julio, and Aris for sharing with me very special moments and interesting talks mostly around a plate of xiaolongbao. And to Cristina for being there all the time, more than a friend she has been a sister.

I am deeply grateful to CENTI Barcelona for their unwavering support, for all their teachings and for being the special people they are. To know them has been a beautiful gift. I also thank to all who were and are no longer, but in its wake left an indelible enriching mark on my life.

Finally, I am forever indebted to my Mother, and to my sisters for their understanding, endless patience and encouragement at the moments when it was more required. I am also grateful to my daughter Silvia for giving me the motivation to finish this thesis.

Abstract

This doctoral work focuses on identifying and understanding the *intents* that motivate a user to perform a search on the Web. To this end, we apply machine learning models that do not require more information than the one provided by the very needs of the users, which in this work are represented by their *queries*. The knowledge and interpretation of this invaluable information, can help search engines to obtain resources especially relevant to users, and thus improve their satisfaction.

By means of *unsupervised learning* techniques, which have been selected according to the context of the problem being solved, we show that is not only possible to identify the user's intents, but that this process can be conducted automatically.

The research conducted in this thesis has involved an *evolutionary* process that starts from the manual analysis of different sets of real user queries from a search engine. The work passes through the proposition of a new classification of user's query intents; the application of different unsupervised learning techniques to identify those intents; up to determine that the user's intents, rather than being considered as an uni–dimensional problem, should be conceived as a composition of several aspects, or dimensions (i.e., as a multi–dimensional problem), that contribute to clarify and to establish what the user's intents are. Furthermore, from this last proposal, we have configured a framework for the on–line identification of the user's query intent. Overall, the results from this research have shown to be effective for the problem of identifying user's query intent.

Resumen

Este trabajo doctoral se enfoca en identificar y entender las *intenciones* que motivan a los usuarios a realizar búsquedas en la Web a través de la aplicación de métodos de aprendizaje automático que no requieren datos adicionales más que las necesidades de información de los mismos usuarios, representadas a través de sus *consultas*. El conocimiento y la interpretación de esta información, de valor incalculable, puede ayudar a los sistemas de búsqueda Web a encontrar recursos particularmente relevantes y así mejorar la satisfacción de sus usuarios.

A través del uso de técnicas de *aprendizaje no supervisado*, las cuales han sido seleccionadas dependiendo del contexto del problema a solucionar, y cuyos resultados han demostrado ser efectivos para cada uno de los problemas planteados, a lo largo de este trabajo se muestra que no solo es posible identificar las intenciones de los usuarios, sino que este es un proceso que se puede llevar a cabo de manera automática

La investigación desarrollada en esta tesis ha implicado un proceso *evolutivo*, el cual inicia con el análisis de la clasificación manual de diferentes conjuntos de consultas que usuarios reales han sometido a un motor de búsqueda. El trabajo pasa a través de la proposición de una nueva clasificación de las intenciones de consulta de usuarios, y el uso de diferentes técnicas de aprendizaje no supervisado para identificar dichas intenciones, llegando hasta establecer que éste no es un problema unidimensional, sino que debería ser considerado como un problema de multiples dimensiones, donde cada una de dichas dimensiones, o facetas, contribuye a clarificar y establecer cuál es la intención del usuario. A partir de este último trabajo, hemos creado un modelo para la identificar la intención del usuario en un escenario *on–line*.

Contents

List of Figures **xiv**

List of Tables **xvi**

1 INTRODUCTION **1**

1.1 Motivation 1

1.2 Contributions 3

1.3 Publications 5

1.4 Organization 7

2 BACKGROUND **9**

2.1 Unsupervised Learning 9

2.1.1 Clustering 10

2.1.2 Association Rules 12

2.1.3 Feature Subset Selection 12

2.2 Related Work 13

2.2.1 From Anomalous State of Knowledge to the User's Intent 13

2.3 Web Usage Mining 15

2.3.1 Sources to Identify the User's Intent 16

2.3.2 How to Determine the User's Intent? 17

3 MODELING INTENT FROM QUERY SESSIONS 23

3.1 Introduction 23

3.1.1 Contributions 24

3.1.2 Outline 25

3.2 User Profiles 25

3.3 Quality of Query Sessions 27

3.4 Document Categories 29

3.5 Data Processing Algorithms 29

3.5.1 Kohonen SOM 30

3.5.2 C4.5 Decision Tree 32

3.6 Data Capture and Preparation 32

3.6.1 Data Mart 33

3.6.2 Data Pre–Processing 34

3.6.3 Data Sampling 37

3.7 Data Pre–Analysis 37

3.8 Unsupervised Results: Clustering of the Data 39

3.8.1 Clustering of the Data in Homogeneous Groups 41

3.8.2 Clusters for Query Sessions: User Types and Quality Profiles 45

3.8.3 Analysis of Resulting Clusters and Subclusters: Documents 47

3.9 Generating a Predictive Model 49

3.9.1 Rule and Tree Induction on Query Dataset 49

3.9.2 Rule and Tree Induction on Document Dataset 52

3.10 Summary 54

4 THE INTENT BEHIND WEB QUERIES 57

4.1 Introduction 57

4.1.1 Contributions 58

4.1.2 Outline 59

4.2 A New Characterization of Web User's Intents 59

4.2.1 User Goals 59
4.2.2 Query Categories 60
4.3 Data Preparation 61
4.3.1 Data Preprocessing 61
4.3.2 Analysis of the Classification of Queries 63
4.4 Data Processing Algorithm 65
4.5 Results 68
4.6 Summary 71

5 MULTI–DIMENSIONAL PERSPECTIVE 75

5.1 Introduction 75
5.1.1 Contributions 78
5.1.2 Outline 78
5.2 Facets for the User Intent 79
5.3 Usefulness of the Facets 82
5.4 Semi–Manual Classification of Queries 84
5.4.1 Query Sampling 84
5.4.2 Clustering the Queries 84
5.4.3 Query Classification Tool 85
5.5 Analysis of the Query Classification 87
5.5.1 Overall Agreement Results 89
5.5.2 Kappa Coefficient Results 89
5.6 Classification Capability of the Facets 90
5.6.1 Best Facet Predictors 91
5.6.2 Association Rules 92
5.7 Summary 95

6 TOWARDS ON–LINE IDENTIFICATION 97

6.1 Introduction 97
6.1.1 Contributions 99
6.1.2 Outline 100

6.2 Dimensions of the User’s Intent 100
6.3 Mathematical Model 100
6.3.1 User Intent Classification 102
6.4 Experimental Setting 108
6.4.1 Dataset 108
6.4.2 Models Tested 110
6.5 Results 110
6.5.1 Impact of WordNet 114
6.5.2 Efficiency and Scalability 115
6.6 Summary 115

7 CONCLUSIONS 117

7.1 Results 118
7.2 Comparison of Techniques 119
7.3 Future Work 121

List of Figures

1.1 Unsupervised Learning models covered in this thesis. 4

3.1 Dataset definition for queries and documents with associated tables of quantiles for selected variables. 35
3.2 Sector diagrams for selected quantiles of key variables in the query (A) and document datasets (B). 40
3.3 Frequencies of document categories in dataset used for clustering. 41
3.4 Kohonen SOM clustering for Queries. 42
3.5 Kohonen SOM clustering for Documents. 43
3.6 Generated rules for users types. 50
3.7 Generated rules for quality categories. 51
3.8 Section of the decision tree generated for document categories. . 53
3.9 High precision decision paths. 54

4.1 Distribution of Queries into *Informational*, *Not-Informational*, and *Ambiguous* user's intents. 64
4.2 Distribution of Queries into Categories. 65
4.3 Percentage distribution of queries into Goals and Categories. . . 66

5.1 Dimensions of user's intent. 76
5.2 Query classification tool. 86

6.1 Example spanning tree (T) on facets. 103

6.2 Factor graph for query facet classification. 105

6.3 The mean (curve) and variance (shown as a vertical bar) of the Hamming error for the overall facet classification for the models *BASE*, *WN*, and *FastQ*. 111

6.4 Number of related words obtained using WordNet for a test query word, and the expected (average) number of related words for a new query word. 112

6.5 Mean (curve) and variance (shown as a vertical bar in each data point) of the number of new words found in the test data, and the number of words for which related words were found in the training data using WordNet with increasing amount of training data. 113

7.1 Paths for future work. 122

List of Tables

3.1 Hypothetical user query session quality profiles. 28

3.2 Hypothetical document categories (ODP + 2 new categories). . . 29

3.3 Quantile ranges for query and document data variables. 36

3.4 Query dataset: Pearson Correlation values for variable *quantiles*. 38

3.5 Document dataset: Pearson Correlation values for variable *quantiles*. 39

3.6 Kohonen clustering of Queries data: averages of input variables for 'level 1' cluster groups. 44

3.7 Kohonen clustering of Document data: averages of input variables for 'level 1' cluster groups. 45

3.8 Average values for key variables used for clustering of document data. 48

3.9 Individual rules evaluated on the testing data. 51

4.1 Manual classification of queries into goals and categories. 67

4.2 Sample of queries with three highest probabilities in the Recreation cluster. 70

4.3 Queries with highest probabilities in the Health cluster. 71

5.1 List of Topics. 80

5.2 Inter-annotation Agreement and Kappa Coefficient for the dimensions of user's intent. 88
5.3 Association rules generated from the query labels. 94

6.1 The dimensions or *facets* which characterize user intent. 101
6.2 Classification accuracy (in %) of the user intent dimensions for the different models. 109
6.3 *FastQ* Classification results for the facet *Task* as *Informational, Not-Informational*, and *Ambiguous.* 114
6.4 Comparison of existing approaches for intent classification with size of the dataset and the classification accuracy. 115

7.1 Unsupervised, Semi–supervised and Supervised models for user's query intent prediction. 120

Chapter 1

Introduction

For everyone who asks receives; the one who seeks finds;
and to the one who knocks, the door will be opened.
Matthew 7:8

1.1 Motivation

Every human interaction as a user on the Web involves accessing an *informational need*. Recognizing the need for information, or as proposed by Belkin [23], that a user has an Anomalous State of Knowledge (ASK), is the starting point that leads people to search for information on the Web. However, as stated by Broder [28], in the Web context the need behind the query is often not informational in nature. In fact, the person may want to interact with the Web or explore a particular Web site which are other type of needs. Due to the inherent difficulty of representing such a *need* through a small set of words –known in this context as a *query*– searching for information is not the simple and straightforward process one would like it to be. From this point of view, a query in an Information Retrieval (IR) system is always regarded as approximate and imperfect [26]. Consequently, there is a gap between a user's information need and the way in which such need is represented. Despite the existence of this gap, an Information Retrieval system should be able to analyze a given query and present the appropriate Web resources that best meet the user's needs.

In order to improve the quality of results, while increasing the user's satisfaction, Web IR systems' current efforts are focused on representing the information needs of their users in a way that not only involves the query but getting as much information related to this need, in an automatic way, transparent to the user and as accurately as possible. From the described scenario, in concordance with the description of next generation of Web search proposed by Baeza–Yates *et al.* [22], emerges the need to determine the goals or the intentions[1] that lead a user to perform a search on the Web. Research advances in this area have addressed the problem of determining the user's intentions as a classification problem, in which starting from a set of queries previously classified (manually) they build predictive models for tagging user's future queries. There are however two main problems in such an approach: the large cost of the manual classification of sets of queries, as well as the subjectivity of people who perform the classification task. In this dissertation, we propose to apply Machine Learning models to identify user's intent. Such a strategy does not suffer from any of the aforementioned problems since we focus on natural identification of user's intent by analyzing the structure of given queries without using any additional information other than the queries themselves.

In a Web search, the user's need, involves an intention. Automatically determine the user's intention is the motivation for the development of this thesis.

The studies presented in this thesis have been developed within the field of Web Mining, specifically Web usage mining. Hence, following its philosophy, we have conducted different analysis of the recorded information in search engines, specifically, in the logs of this kind of systems. Particularly, we have focused our analysis on queries that users submit to Web search engines. Over

[1] Throughout this thesis, the terms *intention* and *intent* will be used to represent the same concept.

these data, we have applied a variety of techniques of unsupervised learning, from the results of which it has been possible to identify the user's intentions, to establish relationships between the various facets involved in the user's intention, as well as to determine behavior patterns of users according to their search intention.

1.2 Contributions

There are four main contributions of this thesis:

1. A framework for identifying a user's query intent. In this work, which is presented in Chapter 4, we propose a new classification of the user's query intent as well as to associate such intent to a set of topical categories. This framework tackles three main aspects: first, a new classification of the underlying user's search intents; second, a deep analysis of a set of queries manually classified; and third, the automatic identification of the users' intents, where the highest precision was 62% which corresponds to the *Informational* intent.

2. The modeling of the user's query intent as a multi-dimensional problem. In this work, detailed in Chapter 5, we introduce the idea that the user's query intent is not a unique dimension, instead we propose that it should be considered as a composition of different dimensions. We propose a set of dimensions extracted from the queries, which we have studied, analyzed and characterized: *genre, objective, specificity, scope, topic, task, authority sensitivity, spatial sensitivity* and *time sensitivity*. Chapter 5 presents

3. The identification of the user's query intent from an unsupervised perspective depending on the application. This implies selecting different unsupervised learning techniques. Although on the Machine Learning literature there is a wide range of *Unsupervised* models, the ones we have selected in this work –see Figure 1.1– both cover the spectrum and, from

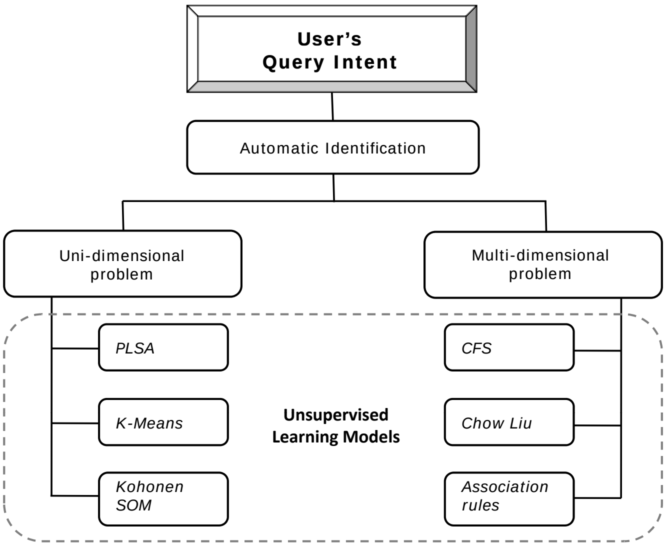


Figure 1.1: Unsupervised Learning models covered in this thesis.

our point of view, are the most representative ones. In all cases we show the effectiveness of Machine Learning models for the problem of user's query intent identification.

4. A model for the on-line identification of the user's query intent. In this work, described in Chapter 6, we pose a generic, extensible framework for inferring the intent of the Web search queries, described in terms of multiple facets. We use a tree structured graphical model for modeling the user intent based based just on the query words. We also incorporate the use of use of WordNet [77] to identify new words present in the search query that are absent in the existing training vocabulary. The average time for query classification is 2 – 3 ms, with an average precision for the task facet of 76%.

The main objective of this thesis is towards identifying and modeling the user's intents through using *unsupervised learning* [38] models, in this work we applied the following techniques, shown in Figure 1.1: *K-means* [69] model used to cluster queries in order to avoid the well known IR problems such as polysemy and synonymy, as well as to reduce the subjectivity of users. The information that a set of related queries offers to a particular one, helps to contextualize a query. *Probabilistic Latent Semantic Analysis (PLSA)* [46] applied to clusters of queries helps in two ways: to determine the user's intentions and the topical categories from a set of queries, and to find the possible relationships that exist among the intentions and the topics. *Kohonen SOM* [62] clustering model is used to cluster and identify user types and profiles of click query documents. Having proposed a set of dimensions involved in the user intent the *Correlation-based Feature Selection* (CFS) [43] model allowed us to determine which are the features that better describe a set of data, and the more valuable features for prediction tasks. *Chow Liu* [70] which is a graphical model together with *WordNet* [77] allowed us to determine and predict the different features that compound the user's query intent. *Association Rules* [102] was utilized for extracting the possible patterns that co–occur in a set of queries.

The data used in this thesis, as well as the proposal of the new classification of the user's query intent, considered both unique and multi–dimensional, are ideas that the author of this thesis share with Cristina González–Caro. Additionally, we give the credits to Vinay Jethava who co-authored the *FastQ* algorithm used in Chapter 6.

1.3 Publications

This work has produced the following publications:

Book Chapters

- David F. Nettleton, **Liliana Calderón-Benavides**, Ricardo Baeza–Yates. *Analysis of Web Search Engine Query Session and Clicked Documents.* In Lecture Notes in Computer Science: Advances in Web Mining and Web Usage Analysis, volume 4811, pages 207 – 226, 2007.

Conference Papers

- Vinay Jethava, **Liliana Calderón-Benavides**, Ricardo Baeza–Yates, Chiranjib Bhattacharyya, Devdatt Dubhashi. *Scalable Multi-dimensional User Intent Identification using Tree Structured Distributions.* To appear in Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, July of 2011.

- Ricardo Baeza-Yates, **Liliana Calderón-Benavides**, Cristina González-Caro. *The Intention Behind Web Queries.* In Proc. of the 13th Symposium on String Processing and Information Retrieval, SPIRE. Lecture Notes in Computer Sciences. Vol. 4209, pages 98–109. 2006.

- David F. Nettleton, **Liliana Calderón-Benavides**, Ricardo Baeza–Yates. *Analysis of Web Search Engine Clicked Documents.* In Proc. of the 4th Latin American Web Congress, LA-WEB'06, Cholula, Mexico. Piscataway: IEEE Computer Society Press, pages 209–219. 2006.

Workshop Papers

- Cristina González-Caro, **Liliana Calderón-Benavides**, Ricardo Baeza-Yates. *Web Queries: the Tip of the Iceberg of the User's Intent.* In Proc. of the User Modeling for Web Applications Workshop at 4th WSDM Conference. Hong Kong, China, 2011.

- **Liliana Calderón-Benavides**, Ricardo Baeza-Yates, Cristina González-Caro. *Towards a Deeper Understanding of the User's Query Intent.* In

Proc. of the Query Representation and Understanding Workshop at 33th SIGIR Conference. Geneva, Switzerland, 2010.

- David F. Nettleton, **Liliana Calderón-Benavides**, Ricardo Baeza–Yates. *Analysis of Web Search Engine Query Sessions*. In Proc. of the Knowledge Discovery on the Web Workshop at 12th ACM KDD Conference. Philadelphia, USA. 2006.

Submitted for Review

- Ricardo Baeza–Yates, **Liliana Calderón-Benavides**, Devdatt Dubhashi, Cristina González-Caro, Libertad Tansini. *A Comparative Study of Machine Learning Techniques Applied to Query Intent Prediction*.

1.4 Organization

The presentation of this thesis is divided into seven Chapters, each of which includes a different perspective that contributes to determine or to model the user's intents. Chapter 2 presents a number of preliminary concepts as well as a review of selected publications related to the topics covered in this thesis. Chapter 3 presents a deep analysis of a Web search query log, in which several features from the user's query session and clicked document are used in order to model the user need. Chapter 4 introduces a new characterization of user's query intents. Chapter 5 proposes a multi–dimensional perspective of the User's Intent. Chapter 6 present a framework for Web query intent identification which is based on the multi–dimensional characterization of user's intents. Finally, Chapter 7 summarizes our contributions and provides guidelines for future work in this area.

Chapter 2

Background

You will know the truth,
and the truth will make you free.
John 8:32

In this Chapter we present a number of concepts related to how to determine the user's intents. As the basis for the development of this work is the use *Unsupervised Learning* models, we first present this concept, and then we give a basic description of the techniques that are used throughout this document. Next, we cover the state of the art that is related to this thesis.

2.1 Unsupervised Learning

Unsupervised learning is a sub–area from Machine learning that studies how systems can learn to represent particular input objects (patterns) in a way that reveals (discovers) the structure of the overall collection of input objects, without any supervision [33]. The obtained representation can be used for reasoning, decision making, predicting things, communicating, etc. [41]. To the light of this work, the obtained representation helps us to determine the user's intent.

In unsupervised learning, clustering is the most used technique. That is the reason why these two terms are used as synonyms. However, there are other techniques such as association rules and techniques that work for feature extraction purposes. The following is a general description of these methods; the details about the used algorithms are further described in the following Chapters.

2.1.1 Clustering

Clustering is a descriptive task where one seeks to identify a finite set of categories or clusters to describe the data [39]. It is the process of organizing objects into groups whose members are similar *in some way*. A cluster is, therefore, a collection of objects which are *similar* between them and are *dissimilar* to the objects belonging to other clusters [13].

The main goal of clustering is to group objects in a way that objects in the same cluster have high similarity to one another, and at the same time are very different from objects in other groups. The greater the similarity within a group and at the same time, the difference with other groups, the greater the quality of the clustering solution will be [96].

There have been many applications of cluster analysis to practical problems. For example, in the context of Information Retrieval, Clustering can be used to group a Web search engine results into different categories. Each category representing a particular aspect of the user's query [79].

An entire collection of clusters is commonly referred to as a *clustering*, and it can be divided into three main categories: *Partition–based, Model–based*, and *Hierarchical* clustering. For the purpose of this work, we applied the two former categories which are described below.

- **Partition–based clustering**

  Sometimes referred to as objective function–based clustering [13]. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained

[32]. While the algorithmic setup is quite appealing and convincing (the optimization problem could be well formalized), one is never sure what type of structure to expect and hence what should be the most suitable form of the objective function. Typically, in this category of clustering techniques, we predefine the number of clusters and proceed with the optimization of an objective function.

The algorithms from this category applied in this thesis are: K–means [68, 94] in Chapters 4 and 5, and Self Organizing Feature Maps (SOM) [62] in Chapter 3. These algorithms were used with the objective to create groups of queries which have in common certain features such as the topic or the user's goal. In particular, the former model, i.e., K–means was used to facilitate and speed up the manual classification of different groups of queries by creating a context to each query, and to give sense to some queries which, from their terms, are incomprehensible, hence difficult to classify.

- **Model–based clustering**

  In model-based clustering, we assume a certain probabilistic model of the data and then estimate its parameters [32]. In this case, it refers to a so-called mixture density model where we assume that the data are a result of a mixture of c sources of data. Each of these sources is treated as a potential cluster.

  The algorithms from this category that were used in this thesis are: Probabilistic Latent Semantic Analysis [46, 58] in Chapter 4, and Chow Liu [31, 70] in Chapter 6. The aim of using these algorithms was to discover the hidden relationships that underlie the queries, and thus, to determine characteristics of the queries such as the topic or the motivation of the user to pose a query.

2.1.2 Association Rules

Association rules mining is another key Unsupervised Learning method that finds interesting associations (relationships, dependencies) in large sets of data items [32]. The items are stored in the form of transactions. Association rules are an essential data mining tool for extracting knowledge from data. The discovery of interesting associations provides a source of information often used by *businesses for decision making*. Some application areas of association rules are market-basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, data preprocessing, etc.

In this thesis we used the Apriori algorithm [29] in Chapter 5 in order to generate association rules. Our purpose was to relate the different facets or dimensions involved in the multi–dimensional representation of the user's intent, as well as to discover possible patterns that co–occur in a set of queries.

2.1.3 Feature Subset Selection

The first problem faced when we are trying to apply machine learning in a practical setting is selecting attributes for the data at hand. To overcome this problem, feature subset selectors are algorithms that attempt to identify and remove as much irrelevant and redundant information as possible prior to learning [43]. The purpose of feature selection is to decide which of the initial (possibly large number) of features to include in the final subset and which to ignore.

Considering the user's intent as a multi–dimensional composition of facets, in Chapter 5 we used the algorithm Correlation-based Feature Selection (CFS) [88] with the purpose to determine which dimensions are the more informative, and should be considered in the task to infer others.

2.2 Related Work

2.2.1 From Anomalous State of Knowledge to the User's Intent

Throughout the history of information search systems, there has been a great effort to establish the aim that leads a user to perform an information search. Research in this area range from defining the information needs of a user as an Anomalous State of Knowledge; passing through determine the behavior of users of Web search engines; up to (recently) establish the intention behind user queries submitted to a Web search engine. These three aspects are defined below.

Anomalous State of Knowledge

In terms of Belkin *et al.* [26] a user information need can be considered as an *anomalous state of knowledge*. The ASK hypothesis proposed by Belkin [23] is that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly. For the purpose of IR, it is more suitable to attempt to describe that ASK than to ask the user to specify her/his need as a request to the system.

Belkin's model can be summarized in three basic steps:

1. A person recognizes a need for information (ASK).
2. That person presents a query (request) to an Information Retrieval system, which returns information in the form of text(s).
3. The person then evaluates the information obtained from the system and determines if his/her need is completely satisfied or not satisfied at all.

ASK can be considered as the starting point for uncountable works trying to determine the user's need, and more recently the user's intent.

User Behavior

Jansen [57] have defined *behavior* as the action or specific goal–driven event with some purpose other than the specific action that is observable. Belkin [25] pointed out that there is a fundamental importance of the users goals, the tasks associated with those goals, with their behaviors, and the intentions underlying the behaviors, and the way they "substantially" affect their judgments of usefulness of the information objects. Still there is little knowledge about the nature, variety and relationships of different information behaviors. In this direction, there has recently emerged a growing interest to discover the user intent as a way to understand their behavior.

User Intent

According to Jansen *et al.* [53] the user intent is the expression of an affective, cognitive, or situational goal in an interaction with a Web search engine. Rather than the goal itself, user intent is concerned with how the goal is expressed because the expression determines what type of resource the user desires in order to address this underlying need [24, 51].

Some classifications of user search intents have been proposed. The first classification that is found in the literature was done by Broder [28], who proposed the following taxonomy of user's intents:

- **Navigational.** This type typically accounts for approx. 30% of all queries. The purpose of such queries is to reach a particular site that the user has in mind, either because he/she visited it in the past or because the user assume that such a site exists.
- **Informational.** This type typically accounts for approx. 40% of all queries. The purpose of this kind of queries is to find information assumed to be available on the Web.
- **Transactional.** This type typically accounts for approx. 30% of all

queries. The purpose of such queries is to reach a site where further interaction will happen. This interaction constitutes the transaction defining these queries.

This taxonomy was further extended by Rose & Levinson [82]. In their work, the authors refined and expanded the transactional query class with a more encompassing resource query class to include viewing, downloading and obtaining resources available on the Web. In spite of the extension by Rose and Levinson, the original categories are the most widely used in the literature, and have been the basis for an important number of studies in the IR area. More recently Nguyen and Kan's [75] proposal of facet classification to further help in possible actions that may be taken by a search engine to aid the search. The facets as defined by them were: Ambiguity, Authority Sensitivity, Temporal Sensitivity and Spatial Sensitivity. The novelty of this classification scheme is that the facets are independent of each other. This means that a query may be Authority Sensitive and Spatial Sensitive at the same time. In their work Hu *et al.* [49] have interpreted the user's query intent through the concepts included in Wikipedia; the authors exemplify their proposal throughout the identification of the following intents: the travel, personal name, and job finding. On the other hand, recent work on query personalization has focused on modeling the query intent [36, 78, 98, 99] and user behavior [37, 101].

2.3 Web Usage Mining

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [93]. One of the main uses of this technique is the understanding of how users behave in the Web, which is the starting point or the basis of the work presented in this document. In particular, we are focused on determining and understanding what the intentions of the user's are when searching for information on the Web.

Research on Web usage mining is mainly based on the analysis of information that is provided implicitly by users and recorded in usage logs, which include search engine query logs and/or website access logs.

Based on the concept of Web usage mining, the following sections describe different features used to represent users as well as the techniques used with the purpose to identify the user's intent.

2.3.1 Sources to Identify the User's Intent

In their work, Kang & Kim [59] used information from document content, links and URLs to classify the user's queries into topic relevance task (informational intent), homepage finding task (navigational intent), and service finding task (transactional intent). Lee *et al.* [65] took into account the past user–click behavior and the anchor link distribution to determine the navigational and informational intents of queries. Liu *et al.* [67] exploited click-through data in order to find navigational and informational/transactional type queries.

Ashkan *et al.* [11, 12] developed a methodology to determine commercial/non-commercial and navigational/informational intent of queries by using ad click-through features, query features, and the content of search engine result pages. Hu *et al.* [49] mapped the user's query intent into the concepts included in Wikipedia, and represented the intents through the articles and categories related to each concept.

Kathuria *et al.* [61] and Jansen *et al.* [52, 53, 51] used several aspects of query sessions, such as the order of a query in a query session, the query length, the result page, and the terms queries, in order to classify queries into the Broder [28] categories of user's intent. Mendoza *et al.* [71, 72] built a decision tree to identify the intent of a user query based on a set of features of the queries and the query sessions. The features they used are: number of terms in the query, number of clicks in the query session, Levenshtein distance, number of sessions with less than n clicks over the total of sessions associated to a query, number of clicks, and PageRank. The authors argument the selection of these features according to their discriminatory capacity.

Rojas *et al.* [45] explored a set of features for the identification of user goals in Web search. In this work the authors proposed the use of different features to represent a query, which they group as: anchor text based features, page content based features, URL based features, query based, and log based features.

Teevan *et al.* [99] investigated the variability in user intent and the relevance search results obtained in terms of implicit relevance measures based on click-through data. They use interdependent sets of features defined using information available in the queries, history of similar queries, and query results.

2.3.2 How to Determine the User's Intent?

The behavior of users, as well as the user's intent, has been studied from different perspectives. The first works in this area were based on statistical analysis of user's queries. This work allowed establishing patterns of user behavior, such as the length of queries, trending topics, and the variation of such trends. Since this kind of analysis is not enough to describe the user, it has emerged the idea to characterize the user's behavior in a way that describes the characteristics of the queries as well as the user's needs. In this regard, the initial approximations were done through manual analysis of queries. Having established a classification of the user's needs, recent works have focused the attention on determining such needs in an automatic way. Below we review the different works in each of the described perspectives.

Statistical Analysis of Web Queries

Analyzes of query logs from Excite [2], done by Jansen *et al.* [56] and Spink *et al.* [92], as well as the analysis done by Silverstein *et al.* [85] using an AltaVista [1] query log, reported various statistics on Web queries. For example, in the year 1999 the mean length of a user query session from the Excite users was 1.6 queries, where the average number of query terms was 2.21. In that same year (i.e., 1999), Altavista users submitted an average 2.02 queries per user session and an average length of 2.35 words per query. Baeza–Yates [15] showed that users from TodoCL search engine [7] use 1.05 words per query.

In their work, Spink & Jansen [90] presented an analysis of different query logs from Excite. This analysis revealed variations in the behavior of users taking into account characteristics that are exogenous to the users, such as the year in which a query was submitted, and the geographic location where the user who submitted the query is situated. With respect to the years, the authors showed that the mean length of queries from the Excite search engine increased steadily from 1.5 in 1996 to 2.6 in 1999; in 2003 the mean length was 2.4. The average number of terms in unique queries was 2.4. In year 1996, the mean query length for users from U.S. and U.K. was 1.5; this value was the same in year 1997 for European users. In year 1999 the average number of words in the queries submitted by U.S. and U.K. users was 2.6, while for users from Europe it was 1.9.

A more comprehensive study was made by Spink *et al.* [91]. In this work, together to some statistics of the queries, the authors presented how the search topics shifted from year 1997 to 2001. In 1997, approximately one in six queries was about adult categories, however by 2001 this amount went down to one in 12, and many of these were related to human sexuality, not pornography. Additionally, by 1999, 83 percent of Web servers contained commercial content and by 2001, Web searching and Web content continued to evolve from an entertainment to a business medium. This analysis was a proof that Internet has changed from being used just for entertainment to be used as an important place to do businesses.

Jansen *et al.* [55] conducted a comparison of nine search engines transaction logs from year 1997 to 2002. Four out of the nine search engines were European, the remaining were from U.S. For this comparison the authors considered features such as: the sessions, queries and the results pages. In terms of sessions, the authors did not find significant differences among the search engines in the number of queries in a session. With respect to the number of terms, although is holding steady, there was a statistical difference in percentage of one-term queries from a German search engine Fireball [4] with respect to the others. The authors argued this fact to the linguistic difference of this search engine. In general terms, Web searchers tend to view fewer documents per Web query. For U.S. search engines the percentage of searchers viewing only the first results page increased from 29% in 1997 to 73% in 2002, while the variability for European searchers ranged from 60% in 1998 to 83% in 2002. This may imply that search engines are improving over the time or that users are dissatisfied sooner.

Manual Analysis of User Needs

Although statistic analysis of Web queries continues to be done, an interest to characterize the behavior of users is now a hot topic in the Information Retrieval community.

Through a user's survey, and an analysis of the manual classification of 400 queries from AltaVista [1] search engine, Broder [28] proposed his categories of user's needs (*i.e.*, *informational, navigational* and *transactional*). Although the amount of queries considered in this study was not significative, this proposal has been the basis for a number of other studies about user's needs and user's intents. Rose and Levinson [82] extended Broder's classification scheme. In their work they tried to establish which classification techniques are effective for this taxonomy. They analyzed a sample of 1,500 queries from Altavista [1], concluding that classification is feasible by examining query logs and click through data. They also noted that ambiguous cases exist, that would be difficult to classify.

Rojas *et al.* [45] conducted an analysis of the manual classification of different sets queries from of TREC collections. The focus of this work was centered on measure the effectivity of using new features to represent the user's goals.

Automatic Identification of Users Needs

Given the vast amount of queries that are daily submitted to a search engine, and the difficulty that represents to work with such amount of queries, a better approach is to characterize the intents of users in an automatic way. Taking the Broder taxonomy of user intent, and with the use of several machine learning techniques, different studies have been done.

Kang & Kim [59] conducted the first automatic classification of Web queries. In this work the authors used a different strategy to meet the need of a user by assigning one task to each query. According to this work, the tasks can be: topic relevance (assumed as informational query), homepage finding (assumed as navigational query) and service finding (assumed as transactional query). Based on the occurrence of query terms patterns in Web pages, Kang and Kim calculated the probability that the class of a user query is the topic relevance or the homepage finding task. As result, Kang & Kim found that URL information and Link information are bad descriptors for the topic relevance task, but good for the homepage finding task.

In the same context, Lee *et al.* [65] carried out an automatic identification of goals (navigational and informational) through the application of heuristics over the clicks on the results pages. The authors proposed two related features, the past user-click behavior and the anchor-link distribution. In this analysis the authors determined that if the number of clicks is near to 1, the goal of the user can be considered as navigational (given that the user had in mind the Web page he/she want to visit); on the other hand, if the amount of clicks is greater than one, the goal could be informational (the user needs to visit some different Web pages before reach a good resource). Although the identification of user goals was done automatically, this study has the limitation that just 50 queries were used, and belong to the Computer Science subject, making this an unrealistic work.

Jansen *et al.* [52, 53] performed an analysis of different samples of queries from seven Web search engines. In this analysis the authors presented a rule–based automated classification system for queries into a hierarchical level of informational, navigational, and transactional user intents. As results, the authors obtained a query classification accuracy of 74%. For the remaining 25% of the queries, the authors argued that user intent is generally vague or multi-faceted, pointing to the need for probabilistic classification. In this work, a clear principle for deriving new rules is not presented and the extension of the rules for a new system of classification or another language is not clear.

Hu *et al.* [49] used the Wikipedia link graph to represent the relations among the concept articles and the relations among articles and categories. Such a graph is traversed randomly to obtain a vector with the probabilities that each concept belong to an intent.

Since the interest to determine the user's intent in Web search is growing rapidly, and we can not argue completeness of the works presented in this state of the art. However, we consider that this sample of works are representative enough of the trends that have been followed the research about user's intent.

Throughout this thesis, we present different works that include the three perspectives for determining the user's intent that were described: *statistical, manual* and *automatic*. As we will see in the following Chapters, we performed several manual analyses of different sets of queries. This classification have been the basis to conduct comprehensive statistical analyses, as well as the reference point to test the behavior of the automatic models used in this work.

As a first approximation to determine the user's intent, the following Chapter presents an analysis of a set of query sessions and clicked documents. In this work we conduct a systematic data mining process to cluster and identify user types and profiles of clicked query documents.

Chapter 3

Modeling Intent from Query Sessions

You ask and do not receive
because you ask wrongly, ...
James 4:3a

3.1 Introduction

Web search log data analysis is a complex data mining problem. This is not essentially due to the data itself, which is not intrinsically complex, and typically comprises of document and query frequencies, hold times, and so on. The complexity arises from the sheer diversity of URL's (documents) which can be found, and from the queries posed by users, many of which are unique. There is also the question of the data volume, which tends to be very large, and requires careful pre-processing and sampling. The analyst may also have the impression that there is a certain random aspect to the searches and corresponding results, given that we are considering a general search engine (TodoCL [7]), as opposed to a specialized search engine (such as Medline [5]) or a search engine contained within a specific website (for example, in a University campus homepage).

Given this scenario, in order to extract meaning from the data, such as user behavior categories, we consider different key elements of the user's activity, such as: (i) the query posed by the user, (ii) the individual documents selected by the user, and (iii) the behavior of the user with respect to the documents presented by the search engine.

Ntoulas *et al.* [76] has evaluated the predictability of Page Rank and other aspects in the Web over different time periods. They found a significant change in the Web over a period of 3 months, affecting page rankings. Baeza–Yates and Castillo [19] trace the user's path through Web site links, relating the user behavior to the connectivity of each site visited. Baeza–Yates *et al.* [21] evaluates different schemes for modeling user behavior, including Markov Chains. Nettleton *et al.* [74] propose different techniques for clustering of queries and their results. Also, Sugiyama *et al.* [95] have evaluated constructing user profiles from past browsing behavior of the user. They required identified users and one day's browsing data. In the current work, the users are anonymous, and we identify behavior in terms of "query sessions". A query session is defined as sequence or queries (on or more queries) made by a user to the search engine, together with the results which were clicked on, and some descriptive variables about the user behavior (which results were clicked on, the time the pages are "held" by the user, etc.).

3.1.1 Contributions

In this Chapter we propose a systematic data mining approach to cluster and identify user types and profiles of clicked documents after a query.

- We define a novel set of *quality* indicators for the query sessions and a set of document categories for the clicked URL's in terms of the most frequent query for each URL. For both cases we analyze the clusters with respect to the defined categories and the input variables, and create a predictive model.

- Our approach has the advantage of not requiring a history log of identifiable users, and defines profiles based on information relating non-unique queries and selected documents/URL's.

3.1.2 Outline

The reminder of this Chapter is organized as follows: In Sections 3.2, 3.3 and 3.4 we respectively present the hypothetical user type, quality profiles and document categories which we propose to identify and predict in the data. In Section 3.5 we describe the data processing algorithms, Kohonen SOM (Section 3.5.1), and C4.5 (Section 3.5.2). In Section 3.6 we describe the data capture and preparation process. In Section 3.7 we describe the data analysis phase, and in Section 3.8 we present the data clustering work. Finally, Section 3.9 describes the predictive modeling with C4.5 rule and tree induction, using the user type and quality profiles as predictive values for query modeling; and the document category labels as predictive values for document modeling. Section 3.10 presents a summary of the Chapter and highlight the most important aspects found in this work.

3.2 User Profiles

We defined as hypothesis, the three main user search intention categories defined by Broder [28], which can be validated from the data analysis. We have to add that this classification is very coarse, therefore the real data does not have to exactly fall into these categories.

Although Broder's categories were detailed in Chapter 2, the following is a description of such categories to the light of the current analysis.

- **Navigational:** this user type typically accounts for approx. 30% of all queries. The user is searching for a specific reference actually known by him, and once he finds it, he goes to that place and abandons the query session. For example, a user searches for *white house*, finds the corresponding URL reference, and then goes to that reference and conducts no further searches. This user would typically use a lower number of clicks and a minimum hold time (the time the user takes to note the reference he is looking for).

- **Informational:** this type typically accounts for approx. 40% of all queries. The user is looking for information about a certain topic, visiting different Web pages before reaching a satisfactory result. For example, a user searches for *digital camera*, finds several references, and checks the prices, specifications, and so on. This user would spend more time browsing (higher document hold time) and would make more document selections (greater number of clicks).

- **Transactional:** this type typically accounts for approx. 30% of all queries. The user wants to do something, such as download a program or a file (mp3, .exe), make a purchase (book, airplane ticket), make a bank transfer, and so on. This user would make few document selections (clicks) but would have a higher hold time (on the selected page). We can confirm the transactional nature by identifying the corresponding document page (for example, an on-line shopping Web page for purchasing a book, a page for downloading a software program, etc.).

In this Chapter, we are interested in applying a methodological data mining approach to the data, in order to identify profiles and rules, which are related to the three main user types defined by Broder [28], and the “session quality” profiles which we will now present in Section 3.3. Also, we inter-relate the two visions of query-session and document, and identify useful features from the overall perspective by analyzing the resulting clusters and profiles. We propose that this is an effective approach for identifying characteristics in high dimensional datasets.

3.3 Quality of Query Sessions

We define four hypothetical categories that indicate query session quality, and which will be validated in Sections 3.7 and 3.8 from the data analysis. We define two categories to indicate a high quality query session, and two categories to indicate a low quality session. The quality of the query sessions can be affected on the one hand by the ability of the user, and on the other hand by the effectiveness of the search engine. The search engine is effective when it selects the best possible documents for a given user query. There are other related issues, such as response time and computational cost, although these aspects are out of the scope of the current study. In the case of user queries, we have chosen some variables which describe the activity: number of search results clicked on; ranking of the search results chosen (clicked) by the user; and duration of time for which the user holds a clicked document. From these variables, we define some initial profiles which can be used to classify (or distinguish) the user sessions in terms of quality. As a first example, we could say that a good quality session would be one where the user clicks on a few documents which have a high ranking (e.g., in the first five results shown), given that it is reasonable (though not definitive) to assume that the ranking of the results is correct with respect to what the user is looking for and has expressed in the corresponding query. With reference to Table 3.1, this profile corresponds to *high1*. Contrastingly, if the user looks a long way down the list of results before clicking on a document, this would imply that the ranking of the results is not so good with respect to the query. Another profile for a good quality query session would be a high hold time, which implies that the user spends a longer time reading/visualizing the clicked document (profile *high2* of Table 3.1).

In the case of low hold times, we cannot assume low quality, because the user may have a *navigational* intent, and therefore finds what he wants and leaves the current session. In the case of an *informational* or *transactional* user's intent, a lower hold time would indicate that the user has not found the content interesting.

Profile (quality of query session)	**high1**	**high2**	**low1**	**low2**
Average hold time of selected documents	–	high	–	low
Ranking of documents chosen	high	–	low/medium	–
Number of clicks	low	–	high	high

Table 3.1: Hypothetical user query session quality profiles.

If we combine this with a high number of clicks, it would indicate that the user has found it necessary to check many results. This profile would correspond to *low2* of Table 3.1. If the user selects many low ranking documents this would also identify that the ordering of the results does not correspond well with the query (profile *low1* of Table 3.1). In order to distinguish the user types in this way, we would need to analyze the content of the documents, which is outside the scope of this study. Therefore, we will limit to quality profiles which can be detected without the need for document content analysis. Table 3.1 summarizes the key variable value combinations together with an indicator of query session quality. Later, we use these profiles to evaluate the query session quality in the clustering results, and as category label for a rule induction predictive model.

The corresponding numerical ranges for the low, medium and high categories were assigned by inspection of the distribution of each variable, together with consultation with the domain expert. The ranges for low, medium and high, respectively, for each of the variables of Table 3.1 are as follows: average hold time of selected documents for a given query, (0-40, 41-60, >60); average number of clicks for a given query, (1-2, 3, >3). In the case of average ranking of documents chosen for a given query, the corresponding labels have an inverse order, that is, high, medium and low, with corresponding ranges of (1-3, 4-5, >5). These ranges are also used for identifying the Broder user search intention categories, as described previously.

ODP Categories	New Categories
Arts, Games, Education, Reference, Shopping, Business, Health, News, Society, Computers, Recreation, Science, Sports, World, Home	Adult, Various

Table 3.2: Hypothetical document categories (ODP + 2 new categories).

3.4 Document Categories

This Section presents the document categories we have used to classify the clicked URLs and which will be validated in Sections 3.8 and 3.9 from the data analysis.

The Open Directory Project defines general categories (see Table 3.2) used to classify search queries. We manually assigned the categories to 1,800 documents, using the most popular query as the classifier. That is, for each document, we identify the most frequent query for that document, and then classify the query using the ODP topics [6]. Therefore, the document is classified in terms of its most frequent query. We defined a new topic Education which substituted Kids and Teens, this latter topic being the original ODP topic. We also defined a class Various for documents which did not seem to classify into any of the other available topics and the class Adult which is not directly defined in ODP.

3.5 Data Processing Algorithms

With the purpose of analyzing the data, in this work we have applied two machine learning models. The first one, Kohonen SOM clustering [62], a non supervised model, with which we created sub-groups of data that share similar features. The second method, C4.5 rule/tree induction [80], is a supervised model from which we create a predictive model for the profiles of user's intents and document categories.

In this Section, we briefly present the algorithm steps and the distance measure for the Kohonen SOM clustering algorithm, and the partition algorithm and criteria used by C4.5 rule/tree induction. They represent two techniques with a completely different approach: the SOM accumulates cases at each lattice node starting with the complete dataset and progressively reducing the local areas (neighborhood) of update; on the other hand, C4.5 starts with a small training subset, testing it on the whole dataset, and progressively increases the size of the subset to include more cases, partitioning the cases based on the values of selected input variables. In general, the Kohonen SOM can be used as a first phase of data mining in order to achieve homogeneous clusters from high dimensional data. Then C4.5 can be used to create classifier models for each cluster created by the SOM. This is confirmed by the results we present in Section 3.9, in which C4.5 produces higher accuracy on individual clusters, and lower accuracy given the whole dataset without clustering as input. Also, the Kohonen SOM presents a *machine learning* solution as an alternative to the traditional statistical approach of K–means clustering [69], often used for clustering term-document data and queries. We could add that the neural network approach of the SOM is adequate for clustering complex datasets with noise and high dimensionality.

3.5.1 Kohonen SOM

The Kohonen SOM is a set of processors which organize themselves in an autonomous manner, only requiring the original inputs and an algorithm to propagate changes in the net. The state of the net resides in the weights (coefficients) assigned to the interconnections between the units. It has two layers: the first layer contains inputs nodes and the second layer contains output nodes. The modifiable weights interconnect the output nodes to the common input nodes, in an extensive manner.

Basic Algorithm

Algorithm 3.1 presents the basic steps of the Kohonen SOM. The global objective is to move the weights towards the cluster centers via the updating of the weights by each input value.

Algorithm 3.1 Kohonen SOM

Step 1: initialize the weight vectors, using random assignments or partially trained weights

Step 2: present inputs to the network.

Step 3: determine the weight vector that is closest to the input vector. Search over complete matrix to find the weight vector with the smallest Euclidean distance difference from the input vector. That is, find i', j' such that

$$\forall_{i,j} \|v - w_{i',j'}\| \leq \|v - w_{i,j}\| \tag{3.5.1}$$

where v is the input vector and i and j range over all the nodes in the matrix.

Step 4: weight adaptation. The adaptation is only applied to weight vectors of nodes within a given neighborhood of the node chosen in Step 3. The neighborhood size is one of the setup parameters, and is gradually reduced during the training run. In this manner, node weights which are further away from the node chosen in Step 3 are modified less. A Gaussian function is then applied to the distance of each node weight vector from the chosen node. That is:

$$w''_{i,j} = w'_{i,j} + \varepsilon \times \exp(\alpha \|v - w'_{i,j}\|^2)(v - w'_{i,j}) \tag{3.5.2}$$

where v is the input vector and the range of i and j is limited to the neighborhood of the node i', j' selected in Step 3. In addition, ε is the “stepsize” and α is a fixed coefficient assigned as the inverse of the neighborhood size.

3.5.2 C4.5 Decision Tree

C4.5 [80] is an induction algorithm which generates rules from subsets (windows) of cases extracted from the complete training set, and evaluates their goodness using criteria based on the precision in classifying the cases. C4.5 is based on the classic method of “divide and conquer” [50]. The main heuristics used are:

- The information value which a rule provides (or tree branch) calculated by $info$ (see below).
- The global improvement that a rule/branch causes, calculated by $gain$ (see below).

Once the training set T has been partitioned in accordance with the n results (outcomes) of a Test X, the forecast for the required information will be the weighted sum of the subsets:

$$info_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \tag{3.5.3}$$

$$gain_X(T) = info(T) - info_X(T) \tag{3.5.4}$$

where T_i represents data subset i and $gain$ measures the information obtained by partitioning T in accordance with Test X. Therefore, the benefit criteria selects a test in order to maximize the information obtained.

3.6 Data Capture and Preparation

In this Section, we describe the original data used, which is organized in a relational data mart structure. We also describe the pre–processing realized to obtain the two datasets, one from the point of view of the user query, and the one from the point of view of the document/URL.

3.6.1 Data Mart

In order to conduct the different tests proposed in this study, we used a set of Web search logs, from the Chilean search engine, TodoCl, captured over a 92 day period from 20th April to 20th July 2004. The data contained in this log file was pre-processed and stored in a relational data base, which enabled us to carry out different analyses on the search behavior of the users of this search engine. From the log file, we have initially selected a sample of 65,282 queries and 39,998 documents.

Before proceeding, we first present some of the concepts used by Baeza–Yates in [21], necessary to understand the data structures used:

- **Query** is a set of one or more keywords that represent a user information need formulated to a search engine.

- **Query instance** is a single query submitted to a search engine in a defined point of time.

- **Query Session** consists of a sequence of query instances by a single user made within a small range of time.

- **Click** is a document selection that belongs to a query session.

- **Document** is an URL Internet address reference.

The data mart we use consists of a series of relational tables which hold transactional and descriptive data about the queries made by the users and the documents clicked by the user from the search results presented to him. The Click table is the most disaggregated of the tables, and contains one line per click by the user. The URL (document) reference is included, together with the time and date of the click, the time the URL was held on screen by the user (hold time), and the ranking of the URL clicked in the list of URL's found. The Query table contains an index to the queries made by the users, including the query terms, number of terms and query frequency. Finally, the QuerySession table links the Query table to the Click table, and aggregates the user sessions from the Click table.

“A priori” and “A Posteriori” Data

Often, in data mining, we consider the descriptive variables in terms of two groups: (i) *a priori*, which are known before an event occurs (such as the launch of a search query) and (ii) *a posteriori*, which are only known after the event has occurred. In the present study, we only have a few *a priori* variables, such as the number of terms and the terms themselves, to describe a user query. On the other hand, we have a significant number of relevant *a posteriori* variables, such as hold times for documents selected, ranking of documents selected, and number of clicks. Therefore, we decided to use both a priori and a posteriori variables in the predictive model of Section 3.9, but calculated exclusively for the given *training* and *test* time periods. That is, the training data used a posteriori variables calculated exclusively from the first 2 months of data, and the test data consists of variables calculated exclusively from the 3rd month of data. This is important for the predictive model of Section 3.9. On the other hand, the unsupervised clustering of Section 3.8 does not have to be restricted to a priori variables, given that it represents the exploration phase.

3.6.2 Data Pre–Processing

The data mart described in Section 3.6.1 was pre-processed in order to produce two datasets: a query dataset, derived from tables Query, QuerySession and Click, and a document dataset, derived principally from the table Click. The resulting data structures are shown in Figure 3.1.

With reference to Figure 3.1, the Query table contains a series of statistics and aggregated values for the queries. *AholdTime* is the average hold time for the URL’s clicked which correspond to the query and *Arank* is the average ranking of those URL’s. *freqQ* is the number of times the query has been used (in the click data), and *numTerms* is the number of terms of the query. *Anumclicks*

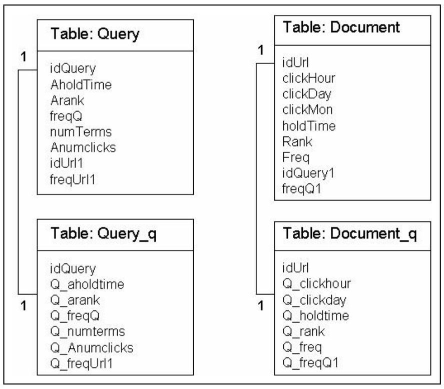


Figure 3.1: Dataset definition for queries and documents with associated tables of quantiles for selected variables.

is the average number of clicks made corresponding to the given query in the click table. Finally, *idUrl1* represents the URL whose frequency was greatest for the corresponding query (from the click data) and *freqUrl1* is its corresponding frequency relative to the given query. These data variables have been selected to create a profile of the query in terms of the statistical data available, which will serve for posterior analysis of the characteristics of search behavior in terms of the queries.

The Document table contains a series of statistics and aggregated values for the URLs referenced in the click table. *clickHour* is the average hour (0 to 23) in which the document was clicked, *clickDay* is the average day (1 to 7, corresponding to Monday - Sunday) on which the document was clicked and *clickMon* is the average month (1 to 3, corresponding to April - July) in which the document was clicked. *holdTime* and *Rank* correspond to the average hold time in the click data and the average ranking of the document in the click data. *Freq* is the number of times the document appears in the click data. Finally, *idQuery1* represents the query whose frequency was greatest for the corresponding URL, and *freqQ1* is its frequency relative to the corresponding URL. These data variables have also been selected to create a profile of the document in terms of the statistical data

Variable	Quantile Ranges
	Query Data
Q_a holdtime	2(0), 3(1-7), 4(8-17), 5(18-29), 6(30-45), 7(46-48), 8(69-105), 9(106-188), 10(189-16303)
Q_arank	1(1), 2(2), 3(3), 4(4), 6(5-6), 7(7), 8(8-9), 9(10-13), 10(14-119)
Q_Anumclicks	2(1), 5(2), 7(3), 8(4), 9(5-6), 10(7-80)
Q_freqQ	3(2), 7(3), 8(4), 9(5-6), 10(7-284)
Q_numterms	1(1), 2(2), 3(3), 4(4-12)
	Document Data
Q_holdtime	1(0), 2(0.14-4.4), 3(4.5-10.5), 4(10.53-18.5), 5(18.57-29.25), 6(29.27-44), 7(44.07-65.91), 8(66-104.8), 9(105-198.5), 10(198.67-40732)
Q_rank	1(1), 2(1.03-1.98), 3(2-2.97), 4(3-3.97), 5(4-4.98), 6(5-6.39), 7(6.4-8.32), 8(8.33-11), 9(11.04-17.25), 10(17.33-200)
Q_clickhour	1(0-11), 2(12-14), 3(15), 4(16-17), 5(18-23)
Q_freq	3(2), 7(3), 8(4), 9(5-6), 10(7-210).
Q_clickday	1(1-2), 2(3-4), 3(5-7)

Table 3.3: Quantile ranges for query and document data variables.

available, which will serve for posterior analysis of the characteristics of search behavior in terms of the documents.

Finally, with reference to Figure 3.1, we observe two additional tables in the lower row, which contain the quantiles of selected variables from the corresponding tables in the upper row. The quantiles have been generated automatically using the SPSS statistical program, and all the variables have been transformed into 10 quantiles, with the following exceptions: *clickhour*, 5 quantiles; *clickday*, 4 quantiles; and *numterms*, 4 quantiles. The number of quantiles was chosen by a previous inspection of the distribution and number of values for each variable. The quantiled versions of the variables were used as inputs to the Kohonen clustering algorithm.

Table 3.3 presents the quantile ranges for query and document data variables. In the case of the variables in each dataset which were not used in the clustering, but were used for cross referencing across datasets, their quantiles were as follows. For the query dataset: Q_freqURL1 2(1), 6(2), 8(3), 9(4), 10(5-166); for the document dataset: Q_freqQ1 3(1), 7(2), 9(3), 10(4-166).

3.6.3 Data Sampling

From the filtered source datasets of queries and documents, by randomly sampling, we created two new datasets of 1,800 records each. These datasets were used as input to the Kohonen SOM. The filtered source datasets had the following number of records: documents 39,998; queries 11,981.

The queries selected must have a frequency greater than 1 (occur more than once) in the click data table. Also the documents (URL's) must also have a frequency greater than 1 in the click data file. The requirement of frequency > 1 for queries and documents, avoids including "one–off" or unique queries and documents in the datasets, as these queries tend not to have any interrelations and create a great dispersion. Finally we filtered records whose hold time was greater than 15 minutes, given that this is a reasonable maximum for normal user sessions.

3.7 Data Pre–Analysis

In this Section we explain the initial analysis of the datasets, using correlation and graphical techniques. We also applied K–means [60] to perform an initial clustering to confirm the initial hypothesis of coherent clusters in the data. In this manner we can identify at this stage any errors in the data or problems due to pre–processing.

With reference to Table 3.4, we can observe a promising degree of correlation between key variable pairs for the complete query dataset comprising of 11,981 different queries. In particular, we can indicate the following correlations (indicated in bold): 0.706 for average hold time with respect to average number of clicks; 0.642 for frequency of the query with respect to frequency of the URL which is recovered most often by this query; and 0.461 for Average ranking of the documents clicked after running the query with respect to average number of

	Avg. hold time	Avg. rank	Query freq.	Num. terms	Avg. num clicks	Freq Url1
Avg. hold time	1.000	**.399**	**.299**	-.061	**.706**	**.309**
Avg. rank	.399	1.000	.188	-.170	**.461**	.049
Query frequency	.299	.188	1.000	**-.233**	.202	**.642**
Num. terms	-.061	-.170	-.223	1.000	-.050	-.173
Avg. num clicks	.706	.461	.202	-.050	1.000	.383
Freq. of Url 1	.309	.049	.642	-.173	.383	1.000

Table 3.4: Query dataset: Pearson Correlation values for variable *quantiles*.

clicks made after running the query.

The document dataset was analyzed in the same way as for the query dataset. In Table 3.5 we observe the degree of correlation between key variable pairs for the complete document dataset consisting of 39,998 different documents. In particular, we can indicate the following correlations: 0.535 for frequency of document with respect to frequency of the most popular query associated with the document; -0.119 for ranking of the document with respect to frequency of the most popular query associated with the document; 0.170 for hold time of the document with respect to frequency of the document; 0.109 for avg. hold time of document with respect to frequency of most popular query associated with the document.

In Figure 3.2 we can see the sector diagrams generated for the quantiles of the key variables in the query (Figure 3.2-A) and document (Figure 3.2-B) datasets.

For the query dataset (Figure 3.2-A) we observe that in the case of *q. freq query* (quantiles of freq. of query), aprox. 55% of the values are in quantile 3, and the second largest proportion is that of quantile 7. For *Avg. number of clicks*, quantile 5 has the largest proportion followed by quantile 2. For the correspondences of quantiles to original value ranges, refer to Section 3.6.2.

For the document dataset (Figure 3.2-B) we observe that in the case of *q. avg. freq. doc* (quantiles of avg. freq. of doc), aprox. 55% of the values are in

	Avg. hold time	Avg. rank	Document freq.	Avg. click hour	Avg. click day	Avg. freq. Query
Avg. Holdtime	1.000	**.086**	**.170**	-.030	.005	**.109**
Avg. Rank	.086	1.000	.020	-.011	.002	**-.119**
Document Freq	.170	.020	1.000	-.043	.062	**.535**
Avg. Clickhour	-.030	-.011	-.043	1.000	.022	-.023
Avg. Clickday	.005	.002	.062	.022	1.000	.032
Avg. Freq. Query	.109	-.119	.535	-.023	.032	1.000

Table 3.5: Document dataset: Pearson Correlation values for variable *quantiles*.

quantile 3, and the second largest proportion is that of quantile 7. For *q. Avg. ranking*, quantile 1 has the largest proportion followed by quantiles 3 and 6, respectively. For the correspondences of quantiles to original value ranges, refer to Section 3.6.2.

In Figure 3.3 we see the frequencies of document categories in the dataset used for clustering. From the total number of 1,800 documents, we observe that 616 (34%) are for category Reference. An example of a Reference document would be: `http://www.navarro.cl/trabajo/datos-historico/mensaje-presidencial-sc-2.htm` and its most frequent associated query is *article 171 of the labor code*. We also see that 217 (12%) of the documents are defined as the Business category. An example of a Business document is `http://www.paritario.cl/lista_comites.htm` and its most frequent associated query is *zanartu ingenieros consultores s.a.*

3.8 Unsupervised Results: Clustering of the Data

In this Section, we explain the clustering process applied to the query dataset using the Kohonen SOM technique, and the following analysis of the data groupings with respect to the user type and quality profiles defined in Section 3.2. The input variables to the query clustering were: *Q_aholdtime, Q_arank, Q_freqQ,*

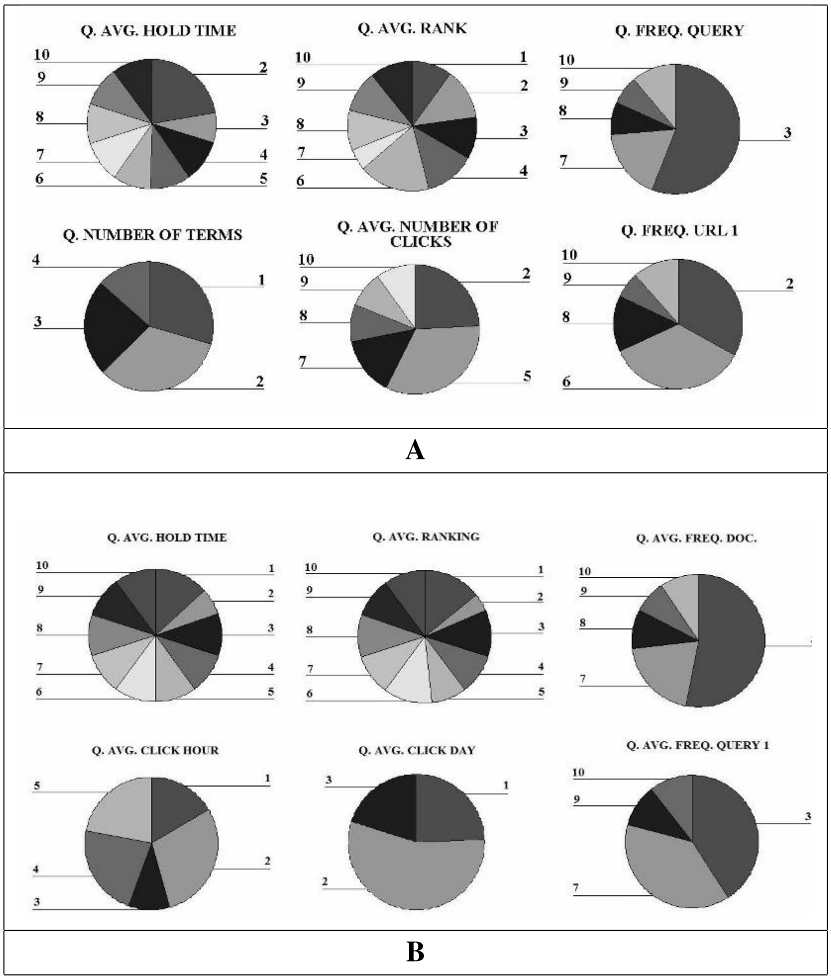


Figure 3.2: Sector diagrams for selected quantiles of key variables in the query (A) and document datasets (B).

Q_numterms and *Q_Anumclicks*. We also detail the clustering process and results for the document data. With respect to the document, the input variables to the clustering were: *Q_clickhour, Q_clickday, Q_holdtime, Q_rank,* and *Q_freq*.

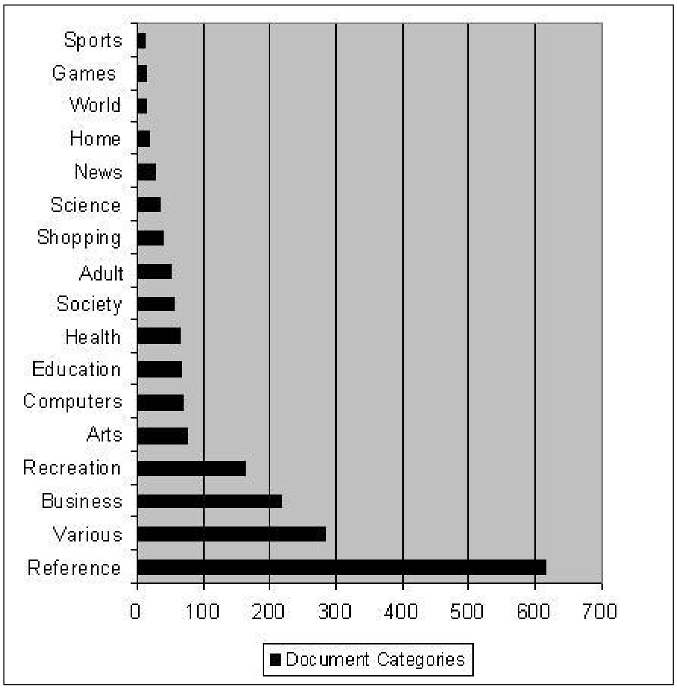


Figure 3.3: Frequencies of document categories in dataset used for clustering.

3.8.1 Clustering of the Data in Homogeneous Groups

The Kohonen SOM algorithm was used as the clustering method for each dataset, using only the quantile values of the selected variables as inputs. The Kohonen SOM was configured to produce an output lattice of 15 x 15, giving 225 clusters, for each dataset. The cluster quality was verified in a post-processing phase, by inspection of activation values and standard deviations. As recommended by Kohonen in [62], we trained the Kohonen net in two stages:

- An ordering stage with a wider neighborhood value, a higher learning rate, and a smaller number of iterations.
- A convergence stage with a smaller neighborhood value, lower learning rate and greater number of iterations.

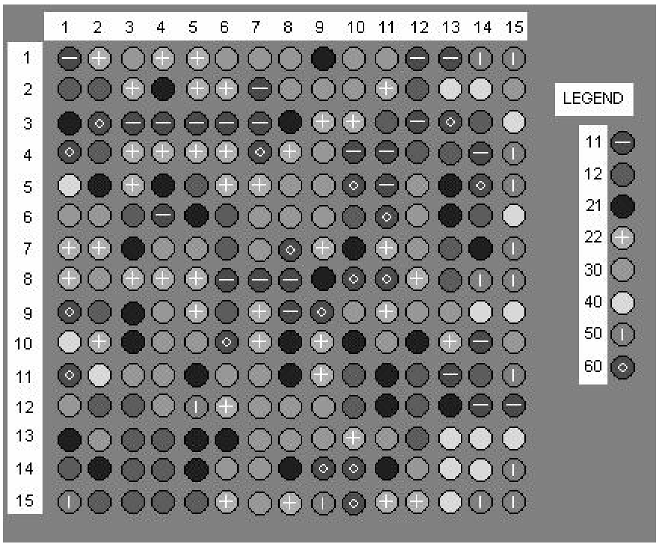


Figure 3.4: Kohonen SOM clustering for Queries.

Figures 3.4 and 3.5 show the graphic output of the Kohonen SOM at the end of the (convergence stage) training run (5,000 iterations) for query data and document data, respectively.

All the training runs were run to complete convergence, that is, when the Kohonen SOM no longer altered the clustering assignments and the patterns (see Figures 3.4 and 3.5) became fixed, which occurred within the 5,000 iterations, for both datasets. In Figure 3.4 (Queries), we observe that the Kohonen SOM has created three major cluster groups, labeled 11, 12 and 30. In terms of the number of cases assigned (as opposed to the number of cluster nodes), by individual inspection we find that the distribution of cases to the query cluster groups of Figure 3.4 is more equal than that of the document cluster groups of Figure 3.5. In Figure 3.5 (Documents), we observe three main groups of clusters, labeled 21, 22 and 30. We also observe four minor cluster groups indicated by labels 11, 12, 40 and 50. Therefore, we state that the Kohonen clustering has identified three major and four minor groupings of clusters. We can now inspect the corre-

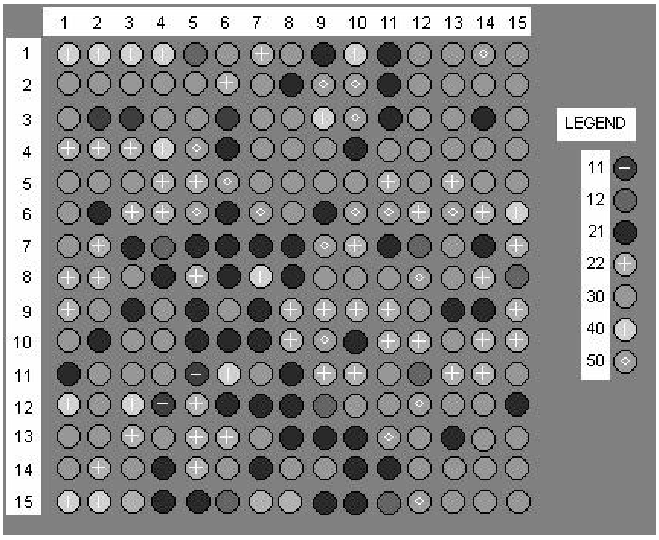


Figure 3.5: Kohonen SOM clustering for Documents.

sponding data in each of these cluster groups, in order to identify distinguishing characteristics in terms of the input variables.

Tables 3.6 and 3.7 list the level 1 cluster groupings generated by the Kohonen SOM 15x15 lattice for the query and documents datasets, respectively. Each of the cluster groups consists of a number of lattice nodes (individual clusters), and these are later detailed in Sections 3.8.2.

Analysis of the Queries Clustering

In Table 3.6, we note that row 1 of the cluster data indicates that cluster group 11 has 191 corresponding queries, and its confidence (activation) value was on average 8.77 with a standard deviation of 2.60. The activation value refers to the neuron activation in the lattice and may be considered as a quality or confidence indicator for the corresponding cluster. With respect to the variable values, we observe that cluster group 11 has the minimum value (2.63) for average query frequency (shown in bold). Cluster group 30 has the maximum values for av-

Queries							**Confidence**	
Cluster Group*	Avg. number of terms	Avg. query freq.	Avg. hold time	Avg. ranking	Avg. number of clicks	Number of Queries	Avg. activation	Stdev. activation
11	3.16	**2.63**	30.87	4.94	1.92	191	8.77	2.60
12	2.24	3.53	103.66	6.86	1.92	214	11.07	2.91
21	1.94	**6.84**	126.59	6.97	2.57	205	11.73	**3.40**
22	2.51	4.59	125.01	5.70	3.28	306	11.86	3.21
30	1.93	4.34	**128.78**	**9.86**	**6.88**	449	14.18	2.82
40	2.04	2.95	**4.16**	**4.42**	1**1.00**	189	**6.44**	2.61
50	**3.45**	2.69	69.24	4.53	1.11	153	7.84	**2.35**
60	**1.53**	4.56	111.03	4.73	2.00	89	9.78	2.97

Table 3.6: Kohonen clustering of Queries data: averages of input variables for 'level 1' cluster groups.

erage ranking of clicked results (9.86), average number of clicks (6.88), and average hold time (128.78). Finally, cluster group 50 has the maximum value for average number of terms (3.45), the second lowest value for average query frequency (2.69) and the second lowest value for average number of clicked results (1.11). We can say that these values are mutually coherent for cluster group 50, given that fewer frequent queries would tend to have a higher number of terms and as they are more specific, the user would click on less of the shown results.

Analysis of the Document Clustering

With reference to the document cluster group data in Table 3.7, we observe that cluster group 12 has the lowest average hold time for documents (34.84), the second lowest value for the average ranking of the documents (6.05), and the highest value for document frequency (6.96), that is, the average number of times that the corresponding documents have been clicked in our dataset. It is coherent that the most popular documents (high frequency) have a high ranking (where position "1" would be the highest ranked), although the low hold time would need further explanation. The low hold time could be due to the corresponding

			Documents (URL's)				Confidence	
Cluster Group*	Avg. hold time	Avg. ranking	Avg. docum. freq	Avg. click hour	Avg. click day	Number of Docs.	Avg. activation	Stdev. activation
11	**241.17**	**14.7**	2.93	**12.33**	3.19	54	10.67	2.78
12	**34.84**	6.05	**6.96**	12.90	2.69	48	10.43	2.34
21	131.29	7.34	5.20	15.46	3.19	408	**11.23**	2.64
22	77.35	**6.01**	4.51	14.37	3.57	330	10.10	**3.08**
30	97.65	7.46	2.70	14.26	**4.32**	759	9.49	2.76
40	118.35	9.11	**2.15**	13.50	1.88	82	**9.64**	**2.33**
50	57.67	9.00	2.39	**18.13**	**1.87**	119	**9.64**	2.29

Table 3.7: Kohonen clustering of Document data: averages of input variables for 'level 1' cluster groups.

pages being of a *navigational* type, therefore the users navigate to them but then disconnect. In the case of cluster group 40, we observe the lowest document frequency (2.15).

In order to cluster the data, the document categories were not given as input. Therefore, given the incidence of greater frequencies of given document categories in specific clusters, we can infer that the input variables given as inputs to the clustering have a relation to the document categories, in the measure that is evidenced.

3.8.2 Clusters for Query Sessions: User Types and Quality Profiles

User types

Now we make some interpretations of the level 1 query clustering results, in terms of the user categories presented in Section 3.2.

- **Navigational.** The query-sessions grouped in level 1 query cluster 40 (see Table 3.6) has a low hold time and a low number of clicks, which has a direct relation with Broder's proposal [28] with respect to the number of documents visited and time spent browsing as a consequence of a query of this type. One example of a *navigational* type query in cluster group 40 is *chilecompra* (in english: *chilepurchase*) with corresponding URL `http://www.chilecompra.cl`, average hold time of 0 and average number of clicks equal to 1. Another example of a typical query in this cluster group is *venta de camisetas de futbol en chile* (in english: *sale of football shirts in chile*) with corresponding URL: `http://www.tumejorcompra.tst.cl/-futbol.php`, average hold time of 0 seconds and average number of clicks equal to 1.

- **Informational:** in query cluster group 30 (see Table 3.6), it can be clearly seen that clusters were generated which grouped the query-sessions whose number of clicks and hold time is high. One example of an informational type query in this cluster group is *cloroplasto* (*chloroplast*) with principal corresponding URL `http://ciencias.ucv.cl/biologia/mod1/-b1m1a007.htm`, average hold time of 731 seconds and average number of clicks equal to 7. Another example is the query *structural engineering software* with principal corresponding URL `http://www.pilleux.cl/-mt771/`, average hold time of 1062 seconds and average number of clicks equal to 8.

- **Transactional:** in query cluster group 12 (see Table 3.6) we can observe medium to high hold times and a low number of clicks, which coincides with our hypothesis for this type of users, although the characteristics are not as strong as for the *navigational* and *informational* user types. Given that we have the queries submitted to the search engine, and we have the documents (Web pages) that the user selected from those retrieved by the search engine, we can confirm individual results as being transactional by visual inspection of the query and of the Web page selected. For example, in cluster group 12, cluster 11,14 we have the following query: *purchase and sale of automobiles* with main URL `http://autos.123.cl/registracion.asp`, with a hold time of 580 seconds and average number of clicks equal to 3. In this case the transaction involves filling in a form.

Quality Profiles

We now interpret the clusters with reference to the session quality profiles presented in Section 3.3 (Table 3.1). *High1*: in all of cluster group 40 we can see a high clicked document ranking (low values) and a low number of clicks (all equal to 1), which corresponds to the hypothetical Profile 1 which indicates high quality. *High2*: cluster 30 has the highest average hold time (see Table 3.6), which is indicative of this quality type. *Low1*: cluster group 30 shows a low/medium clicked document ranking and a high number of clicks, which indicates a problem of low quality according to our definition. On the other hand, we also identified cluster group 30 as having profile *High2*, which is defined in terms of average hold time. This is not necessarily contradictory, given that the queries can show good quality in some aspects, and low quality in other aspects. We would have to investigate the individual level 2 clusters and samples of individual queries and their clicked documents, in order to confirm the problem areas. *Low2*: from the summary statistics of the query clustering, we have not clearly identified this profile among the clusters. We recall that profile *Low2* corresponds to a low quality profile, indicated by a low average hold time, together with a high number of clicks.

3.8.3 Analysis of Resulting Clusters and Subclusters: Documents

Once the Kohonen SOM had generated satisfactory clusters, we selected specific cluster groups by observation, which exhibited potentially interesting data value distributions. For each selected cluster group, we calculated statistics for the corresponding individual cluster nodes of the Kohonen lattice. The selected cluster groups for the documents dataset, cluster groups 12 and 40, are summarized in Table 3.8. This Table presents the average values for key variables used for clus-

	Average values (for each cluster)						Confidence	Query
Cluster	Hold time	Ranking	Freq. doc.	Click hour	Click day	Number of docs.	Avg. activation	Freq. of query 1
Level 2 Document clusters (for level 1 Cluster 12)								
9,12	27.74	**6.11**	**14.7**	10.90	**3.00**	10	11.55	3.10
4,7	**5.93**	5.65	7.57	15.00	2.86	7	10.26	2.86
6,15	58.93	**1.70**	**6.33**	**9.67**	2.83	6	11.14	**4.67**
15,8	**81.47**	4.93	7.14	9.83	**3.00**	6	12.36	3.33
12,7	68.72	5.07	3.00	**17.83**	**2.17**	6	11.11	**2.33**
Level 2 Document clusters (for level 1 Cluster Group 40)								
1,1	**166.25**	**13.55**	2.00	11.90	1.90	10	11.43	1.40
4,4	**11.11**	8.44	2.00	13.56	1.78	9	8.12	1.22
3,12	76.94	**2.61**	2.00	12.33	1.89	9	7.97	1.22
15,6	18.07	0.57	2.00	15.71	1.71	7	5.54	**1.57**
4,1	28.92	21.75	2.00	**10.83**	**2.00**	6	10.7	**1.17**

Table 3.8: Average values for key variables used for clustering of document data.

tering of document data, which correspond to clusters groups in Table 3.6, and one comparative variable not used in clustering (*Freq of query 1*).

With reference to Table 3.8, Cluster Group 12, we observe a relatively high ranking (column 3, 1=highest), and relatively high document frequency (column 4). For Cluster Group 40, we also observe some general characteristic values/tendencies, such as a low document frequency (column 4) and (click day (column 6)). We recall that the days of the week were coded as $1 = Monday, 2 = Tuesday, \ldots, 7 = Sunday$. Individual clusters, such as (6,15) show specific characteristics such as high ranking (1.70), high document frequency (6.33), and low click hour (9.67), which are also primary or secondary characteristics of cluster group 12. With reference to Table 3.8, columns 2 to 8 represent the same variables as those of Table 3.7. In column 1 we see the cluster identifier corresponding to the Kohonen lattice and for the corresponding cluster group, of the first 5 clusters ordered by number of documents assigned. Therefore, in Table 3.8 we observe the summary statistics for individual clusters (9,12; 4,7; 6,15; 15,8; 12,7) assigned to cluster group 12.

Finally, in the last column of Table 3.8, *Freq of query 1* is a variable which was not used as input to the document clustering, and which represents the quantile of the frequency of the query which most coincides with the given document, in the click data.

3.9 Generating a Predictive Model

In this Section we use the C4.5 algorithm to generate a decision tree/ruleset from the two perspectives analyzed in this work: queries and documents.

3.9.1 Rule and Tree Induction on Query Dataset

First we create a model with the user type label defined in Section 3.2 as the classifier category. Secondly, we use the quality label defined in Section 3.3 to create a second predictive model. We train a model on the whole dataset, in order to identify individual rules of high precision which can be useful for user and session classification. The input variables were: number of terms in query, query frequency in the historical data, average hold time for query, average ranking of results selected for query, average number of clicks for query, frequency of the document/URL most retrieved by the query in the historical data, average hour day for submitting of query (0 to 24), and average day for submitting of query (1 to 7).

We note that the original variables have been used as input, not the quantile versions used as input to the Kohonen clustering. This was done to simplify the interpretation of the results in terms of the real data values. We used as training set the first two months of click data, and as the test set we used the third consecutive month. All the statistical variables (averages, sums) were calculated exclusively for the corresponding time periods, in order to guarantee that only *a priori* information was used to train the models. The queries used in the train and test datasets were selected using the following criteria: the same query must occur in the training data (months 1 and 2) and in the testing data (month 3);

frequency of query greater than 1 in the training data, and in the testing data; frequency of the most frequent document corresponding to the query greater than 1, in the training dataset and in the testing dataset. This selection is carried out in order to eliminate "unique" queries and documents, and obtained a total of 1,845 queries for the training and testing datasets to predict user type. In the case of the quality label model, we obtained a total of 1,261 queries for the train and test datasets.

$$\begin{aligned}
&Rule1 : Qholdtime \leq 40\\
&\qquad \Rightarrow class\, navigational\, [69.9\%]\\
&Rule2 : Qholdtime\\
&\qquad\quad Qnumclicks \leq 2\\
&\qquad \Rightarrow class\, transactional\, [50\%]\\
&Rule3 : Qnumclicks > 2\\
&\qquad \Rightarrow class\, informational\, [56.4\%]\\
&Default\, class : navigational
\end{aligned}$$

Figure 3.6: Generated rules for users types.

Figure 3.6 shows the resulting rules induced by C4.5 on the training dataset with the user categories as output. We observe that in order to classify *navigational* type users, C4.5 has used exclusively the hold time, whereas for *transactional* type users C4.5 has used hold time and number of clicks. Finally, in the case of *informational* type users, C4.5 has used exclusively the number of clicks. This is coherent with our hypothetical definitions of these user types: *navigational* users have shorter hold times, *informational* users have many clicks, and *transactional* users have greater hold times and fewer clicks. We could also say that the variables that best differentiate *informational* users from *navigational* users are the hold time and the number of clicks, respectively.

The ruleset was evaluated on the test data (1,845 items), which overall gave 686 errors (37.2%). The accuracy for the individual rules is presented in Table 3.9. We observe that *navigational* is the easiest user type to predict, followed by *informational*, whereas *transactional* seems to be more ambiguous and difficult

Rule	Used	Errors	Label
1	946	285 (30.1%)	navigational
2	228	109 (47.8%)	transactional
3	671	292 (43.5%)	informational

Table 3.9: Individual rules evaluated on the testing data.

to predict.

We also trained a model using the quality categories as the output label. The pruned decision tree generated by C4.5 for the quality categories is shown in Figue 3.7.

```
$Qnumclicks \leq 3$ :
|   $Qrank > 3 \Rightarrow high2$ (171.0)
|   $Qrank \leq 3$ :
|   |   $Qnumclicks \leq 2 \Rightarrow high1$ (523.0)
|   |   $Qnumclicks > 2 \Rightarrow high2$ (25.0)
$Qnumclicks > 3$ :
|   $Qholdtime \leq 40 \Rightarrow low2$ (108.0)
|   $Qholdtime > 40$ :
|   |   $Qrank \leq 3 \Rightarrow high2$ (24.0)
|   |   $Qrank > 3 \Rightarrow low1$ (410.0)
```

Figure 3.7: Generated rules for quality categories.

This tree was tested on 1,261 unseen cases, which gave an overall error of 44%. We observe that *high1* and *low1* are the easiest quality classes to predict, followed by *high2* and *low2* which gave significantly lower predictive accuracies. One possible cause of this could be the range of assignments that we defined in Section 3.3, or due to ambiguities between the different classes.

The overall precision on the whole dataset was not high for the rule induction model, although we did identify several rule *nuggets*, which was our stated objective. One example of a good precision rule was $Qrank \leq 3$ and

$Qnumclicks \leq 2 \Rightarrow$ class *high1* which had only a 31% error on the complete test dataset. It was also found that the user types were easier to predict than the quality classes. One further course of action would be to revise the ranges we assigned to the quality labels in Section 3.3, reassign the labels and rerun C4.5 on the data.

3.9.2 Rule and Tree Induction on Document Dataset

In a similar setting with the queries, for the document dataset we create a model with the document category label (as defined in Section 3.4), as the classifier category. Then we train a model in order to identify individual rules of high precision which can be useful for document and user behavior classification. In this case, the input variables were: avg. click hour, avg. click day, avg. hold time, avg. rank, avg. frequency of the document, and freq. of the most popular query related to the document.

As with the queries, we used as training set the first two months of click data, and as the testing set we used the third consecutive month. The documents used in the training and testing datasets were selected using the following criteria: the same document must occur in the training (months 1 and 2) and in the testing data (month 3); frequency of document greater than 1 in the training data, and in the testing data; frequency of the most frequent query corresponding to the document greater than 1, in the training dataset and in the testing dataset. We ended up with a total of 1,775 documents, which were split in 60/40 for the training and testing datasets, giving 1,058 training cases and 717 test cases. The predictive label (output) was the document category.

Figure 3.8 shows a section of the decision tree induced by C4.5 on the training data, with the document categories as output. This decision path indicates that document category Business corresponds to documents with a hold time greater than 60 seconds and less than 123 seconds, *clickDay* equal to 2 (Tuesday), and ranking equal to or less than 4. There were 27 documents which were correctly classified by this rule from the training data, and 14 doc-

$holdTime > 60:$
$| \quad clickDay = 2:$
$| \quad | \quad holdTime \leq 123:$
$| \quad | \quad | \quad rank \leq 4: \; Business\,(27/14)$

Figure 3.8: Section of the decision tree generated for document categories.

uments were misclassified. This gives a precision for this decision path of $1 - (14/(27+14))100 = 65.85\%$.

By reviewing the complete tree generated, we observe that the C4.5 algorithm has employed all five of the input variables (*holdTime, clickDay, rank, clickHour* and *freq*). We note that *clickDay* was defined as a categorical type variable, whereas the remaining variables were defined as numerical. The most general criteria chosen by C4.5 were *holdTime*, followed by *clickDay*. The most specific criteria used were *clickHour, rank,* and *freq*. We can see extensive use of *clickDay* and *clickHour*, which agrees with other authors who have identified a strong relation between the type of documents selected, the day of the week and the time of day. For example, Business tends to be selected on a weekday $[clickDay = 5]$, and in working hours: $[clickHour \leq 16 : 8(11/5)]$. On the other hand, Home (category 11) tends to be selected at the weekend: $[clickDay = 7 : 11(11/6)]$. This tree was tested on the test dataset consisting of 717 unseen cases, which gave an overall error of 34.9%. We reiterate that the objective of the tree and rule induction was to identify specific higher precision rules, rather than to achieve a high precision global model. Some examples of high precision decision paths are shown in Figure 3.9.

The *decision path 1* correctly classified 11 documents and misclassified 4 documents of category Health, which gives a precision of $1 - (4/(11+4))100 = 73.33\%$. In the *decision path 2* correctly classified 27 documents and misclassified 8 documents of category Reference, which gives a precision of $1 - (8/(27+8))100 = 77.14\%$.

$holdTime > 60:$
$| \quad clickDay = 4:$
$| \quad | \quad rank > 1:$
$| \quad | \quad | \quad holdTime \leq 66: \ Health$
(11/4)

Decision path 1

$holdTime > 60:$
$| \quad clickDay = 3:$
$| \quad | \quad clickHour > 13:$
$| \quad | \quad | \quad rank > 4: \ Reference$
(27/8)

Decision path 2

Figure 3.9: High precision decision paths.

3.10 Summary

In this Chapter we have presented the analysis of a Web search engine query log from two different perspectives: the query session and the clicked document. In the first perspective, that of the query session, we process and analyze Web search engine query and click data for the query session (query + clicked results) conducted by the user. We initially state some hypotheses for possible user types (Section 3.2) and quality profiles for the user session (Section 3.3), based on descriptive variables of the session. In the second perspective, that of the clicked document, we repeat the process from the perspective of the documents (URL's) selected. We also initially define possible document categories and select descriptive variables to define the documents (Section 3.4).

In this Chapter we have contrasted two different techniques, Kohonen SOM clustering (described in Section 3.5.1) and C4.5 rule/tree induction (Section 3.5.2), for mining Web query log data. We have also studied the Web log data from two perspectives: query session and documents. This extends previous results done using other techniques such as K–means. We have detailed all the data mining steps, from initial data preparation (Section 3.6), pre-analysis/inspection (Section 3.7), transformation (quantiles, outliers), sampling, unsupervised and supervised learning algorithms, and analysis of results (Sections 3.8 and 3.9).

The use of machine learning techniques to identify the user categories allows us to confirm the user type “mix” for specific datasets, and to define new user types. In this manner we can classify, on one side, our users and query sessions, and on the other side, the documents and the user behavior with respect to them, in a way which helps us to quantify current user and search engine behavior, enabling us to adapt our system to it, and anticipate future needs.

After this comprehensive analysis, from which we has gained experience working on Web Query Mining, in the following Chapter we introduce a framework for the identification of user’s query intent. In this framework we propose a new categorization of user’s intents and relate such intents with a set of topical categories. Additionally, we used an unsupervised learning model, whose results are contrasted with the manual classification of a set of queries.

Chapter 4

The Intent behind Web Queries

It shall happen that, before they ask, I will answer;
and while they are yet speaking, I will hear.
Isaiah 65:24

4.1 Introduction

Current Web search engines have been designed to offer resources to their users, but with the limitation that the goals or characteristics behind the queries made by them are not generally considered. Given that a query is the representation of a need, a set of factors, in most cases, are implicit within this representation. If we can discover these factors, they can be crucial in the information recommendation process. Techniques such as Web Usage Mining [73] cover the problem to improve the quality of information to users by analyzing Web log data. Particularly, Web Query Mining [20, 16] deals with the study of query logs from data registered in a search engine with the purpose of discovering hidden information about the behavior of users.

There exist several efforts that focus on categorizing the user needs. For example, the categorization proposed by Broder [28] in which, according with the goal of the user, three classes are considered: *Navigational, Informational* and *Transactional*. Broder made a classification of queries through a user survey and manual classification of a query log. This work was later taken up by Rose and Levinson [82], who developed a framework for manual classification of search goals by extending the classes proposed by Broder. In their studies Broder, and Rose & Levinson showed that goals of queries can be identified manually. Lee *et al.* [65] focus on automatic identification of goals (navigational and informational) through the application of heuristics over clicks made by the users on the results offered by the search engine. They proposed two related features, the past user-click behavior and the anchor-link distribution. In the same context, Spereta *et al.* [89] try to establish user profiles by using their search histories; while Baeza–Yates *et al.* [20] discovered groups of related queries through text clustering of documents clicked for the queries, allowing an improvement of the search process.

In general, the approaches presented have tried to make an approximation to the user from different perspectives. However, a model in which the user can be identified by using his/her goals has not been completely developed. From the above, the main goal of this Chapter is to develop a model for identification of the user's interests for a Web search engine using the past interactions stored in the query log. The identification process is made from two perspectives, the first one is from the objectives, or goals, of the users and the second is from the categories in which each of the objectives can be situated.

4.1.1 Contributions

With respect to the state of the art, this work makes the following contributions:

- A new categorization of the user's intents.
- The relationship between intentions and topics.
- The work was done using a set of more than 6,000 real queries, a reference

set two orders of magnitude larger compared to the 50 CS related queries used by Lee *et al.* [65]

4.1.2 Outline

The reminder of this Chapter is organized as follows: In Section 4.2 we present a new characterization of the user's goals (Section 4.2.1), and categories to which a user's query may belong (Section 4.2.2). Section 4.3 contains a description of the query log used in this Chapter, the way this data was preprocessed (Section 4.3.1), and an analysis of the results obtained from the manual classification of the queries in the proposed goals and categories (Section 4.3.2). Section 4.4 describes the PLSA model used in this Chapter to determine the user's intent in an automatic way. In Section 4.5 we present the results obtained from the application of PLSA to determine the user's intent. Finally, Section 4.6 contains a summary of the work presented in this Chapter in which we stress our most important findings.

4.2 A New Characterization of Web User's Intents

In order to determine the user's intentions during the informational search first, we propose a new categorization of the user goals, and second, we relate the goals to a set of topical categories. This information enables us to focus on a closer view of the user when he/she is searching for information on a Web search engine.

4.2.1 User Goals

We define the following three categories of the intents or goals that motivate a user to conduct a search:

- **Informational.** With this type of queries, users exhibit an interest to obtain information available in the Web, without considering the area of knowledge of the desired resource. Examples of *Informational* queries are: *'what is thermodynamics?'*, *'article 150 of the civil code'*, or *'technological advances in veterinary medicine'*.

- **Not-informational.** These queries reflect the interest of the user in obtaining resources or targeting a specific transaction such as buying, downloading, or booking. Some *Not-Informational* queries are:*'Web hosting free'*, *'notebook sales'*, or *'online tv'*.

- **Ambiguous.** These are queries whose intent cannot be inferred directly from the query terms (in some cases because the judge cannot accurately detect the interest category of the user). Some examples of this kind of queries are: *'airbrush'*, *'peeling'* or *'chilean books'*.

4.2.2 Query Categories

A key point in the process of identification of user's query intent is to establish the topic that a query belongs to. The discovery of the kind of information requested, as well as the topical associations of the query allow us to situate the user's query intent in a particular knowledge area, and to link the user to the specific characteristics of such area (or in which he/she wants to be linked).

In order to determine the user's intent, we used the first level categories contained in the Open Directory Project, ODP [6]. The ODP categories are: *Arts, Games, Kids and Teens* (in this work referred to as *Education*), *Reference, Shopping, World, Business, Health, News, Society, Computers, Home, Recreation, Science,* and *Sports.* In addition to these categories we have considered the following three aspects:

- **Various.** For queries that, based on their content, seem to belong to more than one category. Some examples of this kind of queries are: *'human resource consultants'*, *'croaked baked'*, or *'european hazelnut'*.

- **Adult.** This category is added due the large volume of queries of this kind. It is important to remark that although this category does exist in the ODP, it is not explicitly defined in the first level of categories.

- **Other.** For queries which cannot be classified into any of the aforementioned categories. Some queries in this category are: *'176'*, *'counterfeiting'*, or *'roll roads'*.

4.3 Data Preparation

The results reported in this work are based on a log sample from the Chilean Web search engine TodoCL [7]. The sample contains 6,042 queries having clicks in their answers. There are 22,190 clicks registered in the log, and these clicks are over 18,527 different URLs. Thus, on average users clicked 3.67 URLs per query.

4.3.1 Data Preprocessing

In order to have a point of reference that allows validating the results obtained from a machine learning model, the complete set of queries was manually classified. Before carrying out the manual classification, the queries were clustered. With this purpose, each query was represented as a term-weight vector compounded by the terms appearing in the selected Web pages for such query. *Stopwords* (frequent words) were eliminated from the vocabulary considered. Following Baeza–Yates *et al.* [20], each term was weighted according to the number of occurrences and the number of clicks of the documents in which the term appears.

The algorithm used to cluster the queries was K–means [68, 94]. This algorithm follows a partitional approach and is based on the idea that a center point can represent a cluster. K–means use the notion of a centroid, which is the mean or median point to a group of points.

Algorithm 4.2 presents the basic K–means clustering algorithm.

Algorithm 4.2 Basic K–Means

```
1. Select $K$ points as the initial centroids.
2. Assign all points to the closest centroid.
3. Recompute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids do not change.
```

In this work K–means was configured using the following main parameters:

- **Similarity function:** Use to compute the similarity between two documents d_i and d_j. In the context of this work d refers to a query. We selected the commonly used cosine function, which is defined as:

$$cos(d_i, d_j) = \frac{d_i^t d_j}{\|d_i\| \|d_j\|} \tag{4.3.1}$$

- **Criterion function:** The value of a function that is used as criteria to stop the clustering process. This criterion function, see 4.3.2, computes the clustering by finding a solution that separates the documents of each cluster from the entire collection [103]. Specifically, it tries to minimize the cosine between the centroid vector of each cluster and the centroid vector of the entire collection. The contribution of each cluster is weighted proportionally to its size so that larger clusters will weight higher in the overall clustering solution. That is:

$$minimize \sum_{i=1}^{k} n_i \frac{\sum_{v \epsilon S_i, u \epsilon S} sim(v,u)}{\sqrt{\sum_{v,u \epsilon S_i} sim(v,u)}} \tag{4.3.2}$$

where k is the number of clusters, S is the total objects to be clustered (i.e., the queries), S_i is the set of objects assigned to the ith cluster, v and u represent two objects, and $sim(v, u)$ is the similarity between two objects.

- **Number of trials:** number of different clustering solutions to be computed. This parameter was fixed to 10.
- **Number of iterations:** maximum number of refinement iterations to be performed within each clustering step. This parameter was fixed to 10.

In order to cluster the data we used the CLUTO implementation of the K-Means algorithm [60].

Since in a single run of a K–means algorithm the number of clusters k is fixed, we determined the final number of clusters by performing successive runs of the algorithm. By analyzing the results from the different runs of the K–means algorithm, we selected $k = 600$.

Semi–Manual Classification of the Queries

Although the set of queries was manually classified by an expert, we consider this as a *semi–manual classification* given that, through the clustering process implicit information is automatically added to the query. This information is the context that a set of similar queries can give to a single one; however, once the queries were manually classified, the clusters were dissolved, and the queries were considered again as individual items that belong to the general set. In this way we can speedup a direct manual classification by at least one order of magnitude.

4.3.2 Analysis of the Classification of Queries

This Section describes the results obtained from the classification with respect to the user's query intents and the topical categories that were presented in Sections 4.2.1 and 4.2.2, respectively. Additionally, some relations between intents and categories are included.

- **User query intents**. The Figure 4.1 shows the distribution of the queries into the *Informational*, *Not-Informational* and *Ambiguous* user's query intents. The intent with the highest number of queries was *Informational*,

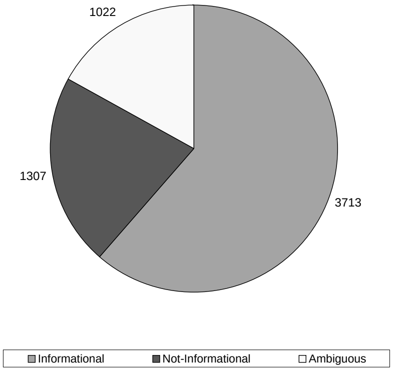


Figure 4.1: Distribution of Queries into *Informational*, *Not-Informational*, and *Ambiguous* user's intents.

61.4% of the queries. Queries in this category are those which are not related to objects such as mp3 files or photographs, names of artists, or with any transaction of products or services on–line. The *Not-Informational* intent have grouped 21.6% of the queries, and the *Ambiguous* intent 17%. Although not all the intents proposed by Broder [28] are the same as the ones we are considering in this work, we agreed with his work, and with some other works based on it, such as [82, 53], in the proportion of *Informational* queries.

- **Query topical categories**. A graphical representation of the manual classification of queries into categories is presented in Figure 4.2. The categories with higher amount of queries are Recreation and Business. This confirms the search behavior of people that has been well described in [92, 54].

- **Relation between intents and categories**. For each of the topical categories, Table 4.1 contains the distribution of queries into the *Informational*, *Not-Informational* and *Ambiguous* intents. Additionally, Figure 4.3 shows

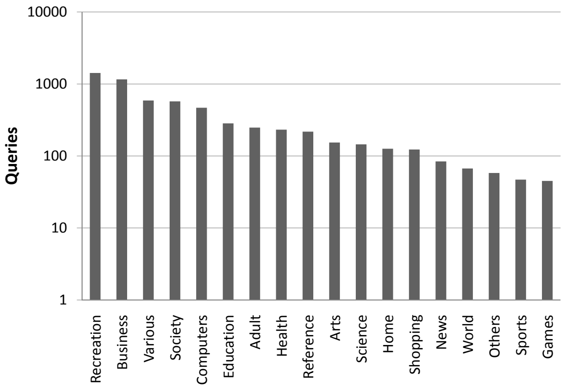


Figure 4.2: Distribution of Queries into Categories.

the percentage distribution of the topical categories into the user's intents. Queries grouped as *Informational* belong to categories such as News, Science, Society, Business, and Education, that are categories in which people are interested in specific resources of information. Queries that were grouped as *Not-Informational* belong to categories such as Recreation, Adult or Games, in which the user's intent is, in most of the cases, to find Web sites offering this kind of resources. Most of the queries grouped as *Ambiguous* fell into the categories Various and Others. This is due to the fact that their terms are not clearly stated, or do not reflect clearly what the user wants, and hence, it is quite difficult to take a decision about how to classify such queries.

4.4 Data Processing Algorithm

To better understand the intents that motivate a user to query a Web search engine, we propose a technique that can automatically characterize the user behavior in order to discover the hidden semantic relationships among his/her queries. Based on these considerations, Probabilistic Latent Semantic Analysis (PLSA)

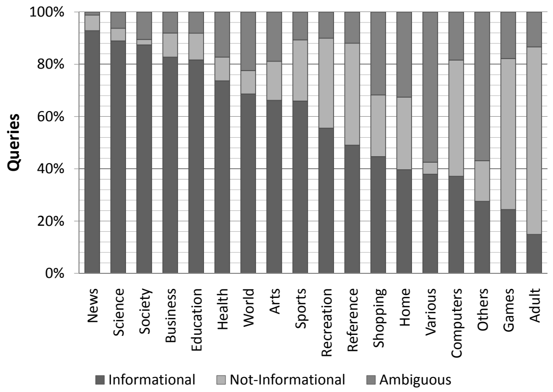


Figure 4.3: Percentage distribution of queries into Goals and Categories.

[46] was the technique selected for this study.

PLSA is a probabilistic variant of Latent Semantic Analysis (LSA) that provides a more solid statistical foundation than standard LSA and has many applications in information retrieval and filtering [84], analysis of Web navigational patterns [66], text learning and related areas [47]. The core of PLSA is a statistical model which has been called the *aspect model* [47]. The aspect model is a latent variable model for general co-occurrence data which associates a hidden (unobserved) factor variable $z \in Z = \{z_1, z_2, \ldots, z_k\}$ with each observation.

The basic idea of the PLSA is to hypothesize the existence of a hidden or *latent* variable z (e.g. a user query intent) that motivates a user to submit a query q using the term w.

In the context of user query intents, an observation corresponds to an event in which a user submits a query q using the word w. The space of observations is represented as an $m \times n$ co-occurrence table (for our case of m queries, i.e., $Q = \{q_1, q_2, \ldots, q_m\}$, and n words, i.e, $W = \{w_1, w_2, \ldots, w_n\}$). The aspect model can be described as a generative model:

- select a query q_i from Q with probability $P(q_i)$

Category	**Informational**	**Not-Informational**	**Ambiguous**	**Total**
Recreation	789	489	142	1,420
Business	960	107	93	1,160
Various	224	27	339	590
Society	501	12	60	573
Computers	174	208	86	46
Education	232	29	23	284
Adult	37	178	33	248
Health	171	21	40	232
Arts	102	23	29	154
Science	129	7	9	145
Home	50	35	41	126
Shopping	55	29	39	123
News	78	5	1	84
World	46	6	15	67
Other	16	9	33	58
Sports	31	11	5	47

Table 4.1: Manual classification of queries into goals and categories.

- given this query q_i, pick a latent factor z with probability $P(z|q_i)$
- given this latent factor z and the query q_i, generate a data item w_j from W with probability $P(w_j|z)$

As a result we obtain an observed pair (q_i, w_j), while the latent variable z_k is discarded. Translating this process into a joint probability model results in the following expression:

$$P(q_i, w_j) = P(q_i)P(w_j|q_i) \tag{4.4.3}$$

$$P(w_j|q_i) = \sum_{z \in Z} P(w_j|z)P(z|q_i) \tag{4.4.4}$$

This model is based on two independence assumptions: first, the observations pairs (q_i, w_j) are assumed to be generated independently; second, conditioned on the latent factor z, the word w_j is assumed to be generated independently of the specified item q_i. Using Bayes rule, we can rewrite Equation 4.4.3 into:

$$P(q_i, w_j) = \sum_{z \in Z} P(z)P(q_i|z)P(w_j|z) \tag{4.4.5}$$

PLSA uses the Expectation-Maximization (EM) [35] algorithm to estimate the probability values for $P(z)$, $P(q_i|z)$ and $P(w_j|z)$, which measure the relationship between the hidden variable and the observations (Q, W), by choosing them such that the likelihood $L(Q, W)$ is maximized.

The algorithm alternates two steps:

- An expectation (E) step, where posterior probabilities are computed for the latent variable, based on the current estimation of the parameters.

- A maximization (M) step, where parameters are re-estimated to maximize the expectation of the complete data-likelihood.

The computational complexity of this algorithm is $O(mn\ell)$, where m is the number of queries, n is the number of words, and ℓ is the number of factors. Since the observation matrix is, in general, sparse, the memory requirements can be dramatically reduced by using an efficient sparse matrix representation of the data.

The implementation of this model was taken from PennAspect [84], a well tested software tool [83] for information filtering and retrieval.

4.5 Results

Before analyzing the results obtained through PLSA, and having in mind that this work is focused on finding query user intents, the categories in which these intents can fall, as well as the possible relationships among query user intents and categories, it is important to consider that in a common process of information search we are exposed to factors such as the following:

- The lack of precision between the transformation of a user information need, and the set of key terms used to represent it [23].
- The lack of accuracy from search engines to provide an answer including aspects such as imprecision, vagueness, partial truth, and approximation [14].

In spite of these factors, taking advantage of the results offered by PLSA, and its capability to provide information about the degree of membership of each query to each cluster generated, we can make an analysis of the content of each cluster. This information gives us the possibility to detect direct and indirect relationships between queries, leading us to discover interesting information about the intents of users.

From the results, we found that approximately 73% of the queries grouped as *Ambiguous* belong to categories such as Adult and Entertainment. It is important to notice that none of the queries manually classified as Health are part of the *Ambiguous* query user intent. The reason for this is that a person usually has in mind the name of an specific disease or medicine. From the queries grouped as *Informational*, about 76% are related to References, Education, Health, Computers, Society and Home. On the other hand, from the queries grouped as *Not-Informational*, about 70% were labeled (in the manual classification) as Computers, Entertainment, Society and Adult. The difference between queries that belong to the *Ambiguous* and *Not-Informational* clusters is that the content of the later one is mainly about photos of people from the show business, parts of computers, and downloads of software and music; while in the case of the *Ambiguous* cluster the query terms do not reflect a specific need.

From the point of view of categories, in order to determine the relationships between the different queries we considered the topics described in Section 4.2.2

Query	Prob1	CId	Prob2	CId	Prob3	CId
los jaivas main works	1.76E-03	6	1.99E-09	0	4.30E-12	7
ricardo arjona songs	1.51E-03	6	2.01E-08	7	2.12E-42	11
madonna erotic	1.50E-03	6	1.83E-08	7	2.20E-42	11
porto seguro cd	1.50E-03	6	2.68E-08	7	1.29E-43	11
rata blanca songs	1.50E-03	6	2.69E-08	7	8.84E-43	11

Table 4.2: Sample of queries with three highest probabilities in the Recreation cluster.

as the hidden variables which motivate a user to submit a query.

One of the most important aspects to highlight, is that despite the fact that the number of clusters used to make the clustering process for categories identification was taken from the high–level ODP categories plus the three categories that we have introduced (i.e., 18 categories), PLSA was not able to create a cluster for each of these categories. The results show a strong grouping of queries around some categories such as Recreation, Business, Society, Computers, Education, Adult, and Health. The model could not create groups for categories such as Arts, Sports, Science or Games. The possible reasons for this are: the small number of queries in these categories (see Section 4.3.2), and that these categories can be treated as a subset of other stronger categories.. This information allows us to:

- Ratify that most of the categories used in the manual classification are clearly defined. However, there are other categories which are difficult to distinguish, because they have overlapping content. In contrast, we found other 'possible' categories that we had not considered. These new categories are: *cars* and *law*.

- From this information we can identify existing relationships among categories. Table 4.2 shows an example of these relationships; the Table contains a sample of queries belonging to the cluster 6, which grouped

Query	Probability
electroconvulsive therapy	1.75E-03
nasal polyps	1.53E-03
dental hygienist	1.51E-03
hepatitis	1.41E-03
viagra	1.03E-03

Table 4.3: Queries with highest probabilities in the Health cluster.

queries related to Recreation or Entertainment. By observing the probability of each one of these queries (Prob1, Prob2, Prob3), to belong to each cluster (CId), the highest values are for clusters labeled as Business (cluster 7) and Adult (cluster 11).

In general terms it is reasonable to think that queries that were grouped in the Adult category have a high probability to belong to the Recreation category. On the other hand, due to the fact that the content of pages answering queries about adult content and entertainment includes terms related to selling or payments of these kind of services, these same queries can be considered as belonging to the Business category.

A particular case was presented by the cluster that grouped queries related to Health. About 70% of queries belonging to this cluster made reference to medicines, diseases or treatment of diseases. The reason for this case is that the medical vocabulary and the terms used to write this kind of queries is very specific, and it is unusual to find problems of synonymy or polysemy. Some examples of queries this cluster can be seen in Table 4.3.

4.6 Summary

In this Chapter we present a first step for identifying user's intents from a Web search engine's query log. We propose that the user's intent can be: *Informa-*

tional, Not–Informational and *Ambiguous* (Section 4.2.1 presents a description of them). Along to this characterization, in this work we analyze a set of topical categories (Section 4.2.2), in which user's intent may be classified. We also make a comprehensive analysis of a set of queries that were manually classified into the user's intent and topical categories (see Section 4.3.2), and relate these two aspects.

In order to identify the user's intents, we applied different techniques. First, a manual classification of a set of queries (see Section 4.3.2). In order to reduce the ambiguity of some queries, as well as to lighten the process, a step in the manual classification was the application of K–Means algorithm to clustering the queries. Later, we carried out an automatic identification of the user's intents by applying the Probabilistic Latent Semantic Analysis model (presented in Section 4.4).

From the application of PLSA we found that the user's needs which are related to entertainment, adult content or business may be very well detected (this analysis is presented in Section 4.5). Additionally, the results reflected very important relationships between these categories. It is highly probably that a query that was classified in one of these categories, may be classified as one of the other two categories.

There are some categories that were not determined by PLSA. A reason for this is that the terms used to submit some queries, as well as the content of the clicked documents (i.e. the pages that answer to such queries), are general terms which may be used to describe different topics. In this work, eleven out of the eighteen categories proposed were recognized by the unsupervised learning model.

Finally, we found two new –and well defined– categories, which had not been considered. These categories are: *cars* and *law*. This indicates that the ODP categories might be revised to avoid any information overlap.

Through the use of two *Unsupervised Learning* models: K–means and PLSA, in this Chapter we have shown the feasibility to determine the user's intents in an automatic way, without the need of adding more information than the queries. However, from the query terms we consider that we can extract valuable information which can be used to gain more insight about the user's intent. Considering this, in the following Chapter we introduce a different way to represent the user's intent based on the information extracted from the queries. Hence, instead of conceive the user's intent as a single dimension, we propose that it is compound of multiple dimensions.

Chapter 5

Multi–Dimensional Perspective

But the fruit of the Spirit is:
Love, Joy, Peace, Forbearance, Kindness,
Goodness, Faithfulness, Gentleness and Self-Control.
Against such things there is no law.
Galatians 5:22-23

5.1 Introduction

Users who search for information on the Web may have very different intents: shopping, researching on a topic, planning a trip, looking for a quick reference, and so on. Characterizing and identifying the intent of the user queries is one of the most important challenges of modern information retrieval systems. Studies characterizing the user's intent have conceived this intent as a unique dimension [28, 18]. However, such a simple model is often inadequate for capturing the complexity of information seeking in the real world, and search engines need a better understanding of user intent and its multi-dimensional nature. Hence, with the aim to provide a more comprehensive understanding of the user and his/her

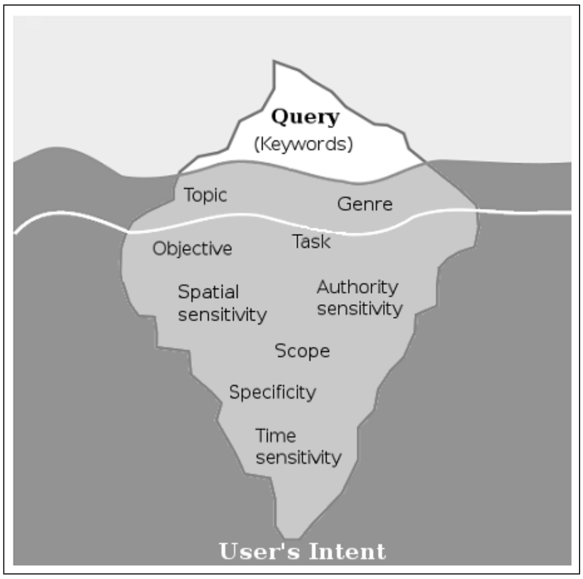


Figure 5.1: Dimensions of user's intent.

intents, in this Chapter we study and evaluate a set of facets or dimensions[1] that give a more accurate representation of users needs and intents. We also show the feasibility of prediction for the dimensions.

Searching the Web for useful resources is not an easy task. Most of the users use short queries to describe their information needs. Starting from the query words, a search engine should be able to accurately infer the user's intents and to guide the user to obtain the actual desired information. In this context, a query can be considered the *tip of the iceberg* of the need for user information, as represented in Figure 5.1, which only shows –in a direct way– a small part of its content, while hiding an important proportion of its potential information.

In order to improve the quality of results, while increasing the users satisfaction, Web IR systems current efforts are focused on representing the information needs of its users in a way that not only involves the query but to get as much

[1] In this Chapter, the terms *facet* and *dimension* will be used to represent the same concept.

information related to this need, in an automatic way, transparent to the user and as accurately as possible. From this arises the need to determine the intents that lead a user to perform a search on the Web. Search Engines require the use of different strategies to meet the user's intents. It is necessary to combine several factors in order to gain insight into the real goal of users' search. For example, is the user searching for a document to read it? or does he/she want to perform a transaction such as buying a product?. Is it important that the retrieved information comes from trusted Web sites? –or the source of the information is not a relevant feature, as long as the content is good–, how important is the information recency? what kind of resource is the user searching for? what the user wants to do with such resource? –download it, read it on–line, or just find a reference–. These questions help us to build a picture of the multi–dimensional nature of the user's intent, and show why is it necessary to place the user's requirements in a wider context.

Several authors have studied Web query logs to explore the issue of defining and discovering user intent in information search on the Web. Broder [28] defined a classification of Web queries into three categories: *informational*, *transactional*, and *navigational*. Rose and Levinson [82] introduced a goal hierarchy and presented manual classification results on $1,500$ queries from Altavista query logs. Following Broder's taxonomy, several authors have focused their work on the automatic classification and characterization of the user intents [53, 65, 67]. Other researchers have worked on large manually annotated data sets, such as the work shown in Chapter 4. Nguyen *et al.* [75] proposed a facet classification to further help in possible actions that may be taken by a search engine to aid the search process. The facets defined by them were: Ambiguity, Authority Sensitivity, Temporal Sensitivity and Spatial Sensitivity. Recently the interest in determining user intentions has spread to commercial [12] and geographical [40] applications.

Although previous studies –from different perspectives– have addressed, the issue of determining user intents, there is still not an approximation that integrates such perspectives. Each study analyzes its own proposal, and albeit the conclusions are useful and the relevance of the results is shown, a comprehensive study that explores the integration and the relation among them is still needed.

The purpose of this work is to delineate the relationships and dependencies that exist among the dimensions involved in establishing the user's intention. Furthermore, we aim to identify the contributions of each dimension in the task of accurately recognizing user's intentions in Web search.

The selection of the facets was done by means of the observation of a set of queries, the characteristics of such queries and the main features that are noticeable in the whole set, or in subsets of queries. Hence, although this is a high number of facets, in this work we wish to ascertain which facets can be considered to be effective descriptors of the user's information needs, and using this information to filter out the facets leaving the most representative ones. The studied facets are: *genre, objective, specificity, scope, topic, task, authority sensitivity, spatial sensitivity* and *time sensitivity*. This research introduces the first four facets, and analyze them together with others proposed by different authors [18, 28, 75].

5.1.1 Contributions

The most important contributions of this Chapter are:

- A new vision of the user's search intent as a problem of multi-dimensional nature.
- This study analyzes nine dimensions involved in the user need, four of which are introduced here.

5.1.2 Outline

The reminder of this Chapter is organized as follows: In Section 5.2 we describe the dimensions we are proposing together with the values they can take. Section 5.3 presents some possible applications of facets from a search engine's perspective. Section 5.4 describes the data we used and the semi-manual classification of queries. Section 5.5 presents an analysis of the manual classification of the queries in terms of the inter-annotation agreement (Section 5.5.1), and the Kappa Coefficient values (Section 5.5.2). In Section 5.6 we make an analysis of the classification capabilities of the facets. This analysis includes determining which the Best Facets Predictors are (Section 5.6.1), as well as the generation of some Association Rules (Section 5.6.2). Finally, in Section 5.7 we present a summary of the Chapter and a discussion of our findings.

5.2 Facets for the User Intent

In this Section we analyze a set of dimensions that can be extracted from a user's query and that should be considered in order to determine the user's intents. Each of the dimensions covers a specific part of the user need, where this need is seen as a composition of a wide spectrum of facets. Below, we describe the dimensions, the possible values they can take, as well as outline some possible practical applications of them.

Genre. This is the broadest dimension from the considered set. *Genre* provides a generic context to the user's query intent, and can be thought of as a meta-dimension. The values that *Genre* can take are:

- *News* when the user interest is related to current events or people (e.g., *Palestinian conflict, catholic university hunger's strike*).
- *Business* are those queries that reflect a need for making a transaction such as a purchase, sale, or lease of a good or service (e.g., *sale of furniture, Internet housing*).
- *Reference* are those queries related to seeking information about a data item, a name, a location, or a resource, with the purpose to get the picture of it (e.g., *history of the Chilean air force, contraceptive method*).

	Topics	
Adult	*Finance*	*Science & Mathematics*
Arts & Culture	*Food & Drink*	*Social Science*
Beauty & Style	*Health*	*Sports*
Cars & Transportation	*Home & Garden*	*Technology & Electronic*
Charity	*Industrial Goods & Services*	*Travel*
Computers & Internet	*Music & Games*	*Undefined*
Education	*Politics & Government*	*Work*
Entertainment	*Religion & belief systems*	

Table 5.1: List of Topics.

- *Community* queries are those related to society, lifestyle, celebrities, Web social groups and interaction tools (e.g., *video chat*, *Ricky Martin*).

Topic. As reported by [42] to truly understand the user query and the quality of the results, a search engine also needs to understand the topical associations to a query. The list of topics used in this work (see Table 5.1) was built from the first level of categories offered by ODP [6], Yahoo! [10], and Wikipedia [9], but deleting categories that are really a *genre* and not a *topic* such as *News* or *Reference*.

Task. We consider *task* [18] as the primary need that is reflected by the query. The possible values for this facet are the following:

- *Informational* when the query reflects an interest for obtaining information from the Web, irrespectively from the knowledge area of the resource to be retrieved (e.g., *what is thermodynamics*, *article 150 of the civil code*).
- *Not Informational* if the interest is towards obtaining resources, or is targeting a specific transaction, such as buy, download, or booking

(e.g., *Web hosting free*, *notebook sales*, *online TV*).

- *Ambiguous* queries means that their task can not be determined and may be either obtaining information or performing an activity (e.g., *airbrush*, *peeling*). To determine the correct intent for this kind of queries we need to know the context of it.

Objective. Represents if the query is aimed to obtain a *Resource* or to do some *Action*. The two values for this facet are:

- *Resource* for us, it means the need for information which is presented in formats such as HTML, PDF documents, videos or slides (e.g., *street map of Santiago*, *Pablo Neruda poems*).
- *Action* describes the need to carry out an activity such as filling a form or subscribing to a mailing list (e.g., *rent of apartments*, *download MSN 5.0*).

Specificity. This dimension is about how specialized is a query. The following are the values that this facet can assume:

- *Specific* if the query contains a name, date, place, an acronym, or even an URL (e.g., *university of Santiago de Chile*, *OAS*),
- *Medium* if the query does not contain neither a name, place, any of the aforementioned items, nor it is very general (e.g., *private universities*, *agricultural machinery*).
- *Broad* if the query is expressed using a single and very general term. The average length of these queries is one (e.g., *universities*, *maps*).

The possible values for the following facets (i.e., *Scope, Authority Sensitivity, Spatial Sensitivity*, and *Time Sensitivity*), are: *Yes* or *No*.

Scope. The scope aims at capturing whether the query contains polysemic words or not. A positive *scope* will hint to the possibility that the user is looking for certain information issuing a query whose words have different meanings (e.g., *wood*, *dove*).

Authority Sensitivity. Through this facet, it is possible to determining whether the query is designed to retrieve authoritative and trusted answers [75], (e.g., *clinical hospital of the university of Chile*, *hotmail.com*). According to Broder's taxonomy, *navigational* queries are sensitive to authority.

Spatial Sensitivity. This dimension reflects the interest of the user to get a resource related to an explicit spatial location, such a city or a place (e.g., *hotels in Valparaiso*), or to find a location that is not mentioned, but that might be inferred from the query terms [75, 22], (e.g., *nearest hotel*, *restaurant*).

Time Sensitivity. This dimension captures the fact that some queries require different results when posed at different times [75], (e.g., *US president* vs. *US president 2009*).

An important aspect to consider is that we can describe classes of queries through composite facets. For example, most of Transactional queries can be defined as *genre=Business* together with *task=Not Informational*.

5.3 Usefulness of the Facets

Although it would be impossible to claim facet completeness, search engines could take advantage of the knowledge of these facets. Below, we delineate some useful applications of these facets, and the way a search engine can benefit from them.

Genre. This facet clearly delimits the search engines area of search, and is expected to make the search faster and more accurate. For example if the *genre* is *Community*, the pages that have to be looked at are blogs, forums, chats, and social networks, such as Facebook [3].

Topic. Although this is one of the first features that are considered by the search engines, they should ideally present the answers to the user organized by *topic*.

Task. If the query is *Informational*, then Web resources to be recommended should avoid those with transactional content, or the ones where the user has to assume an active role of interaction, such as fulfilling a form.

Objective. If the user wants to perform an *Action*, then he/she may also be interested in commercial sites that offer a product or service, hence appropriate ads may be presented to the user.

Specificity. This relates to diversity in Web search engine results: for ambiguous or broad queries, Web search engines may try to diversify the results by trying to cover many aspects/facets of the query in the top 10 results, so that more users can actually find what they are looking for.

Scope. The answers presented to the user by the search engine should ideally be presented by *topic*, and trying to cover most of the topics related to the polysemic words, allowing the user to select the right one.

Authority Sensitivity. For this dimension the search engines have the opportunity to be very selective regarding the authority of the Web pages presented to the user. For example, it is critical to trust the pages from where to download anti–virus software, and it is essential to reach the right page for all the *navigational* queries.

Spatial Sensitivity. Although a query does not mention a place, search engines should be able to identify the *spatial sensitivity* to offer information related to the local area from which the query was submitted, emphasizing in queries that reflect a commercial need. For example: *school*, *tai chi classes*, or *cars selling*.

Time Sensitivity. Search engines can take advantage of the *time sensitivity* of queries to suggest related information that occurred in the same period as what it is being sought for. Furthermore, this information can be used to recommend resources following a temporal order, as it is done with news.

5.4 Semi–Manual Classification of Queries

In this Section we describe the data used in this work, as well as the semi–manual classification of the queries. In the same way that in Chapter 4, in this work we say that this is a *semi-manual* classification of the queries considering that through the clustering process, a query is enriched with the information of a group of queries related to it.

5.4.1 Query Sampling

We processed and manually classified a sample of 5,249 queries from the TodoCL search engine query log. The original data set belongs to a log file that registered the activity of the search engine for a period of one year. The data set contains 252,675 unique queries related to 365,737 query sessions. In the original log were registered 1,453,873 selections (clicks) over a set of 451,647 different URLs.

As a way to work with queries with different level of popularity (in Chapter 4 only popular queries were considered), the set of queries was randomly sampled after calculating their popularity to capture the Zipf's law distribution: 15% were taken from to the set of the most popular queries, 15% from the (long) tail set, and the remaining 70% was sampled from the middle set (i.e., queries with average popularity).

5.4.2 Clustering the Queries

In a similar way as described in Chapter 4, before carrying out the manual classification, the queries were clustered using the K–means algorithm. The details about the algorithm and the clustering process are mostly the same, however in this work the number of clusters to generate was fixed to $k = 300$.

5.4.3 Query Classification Tool

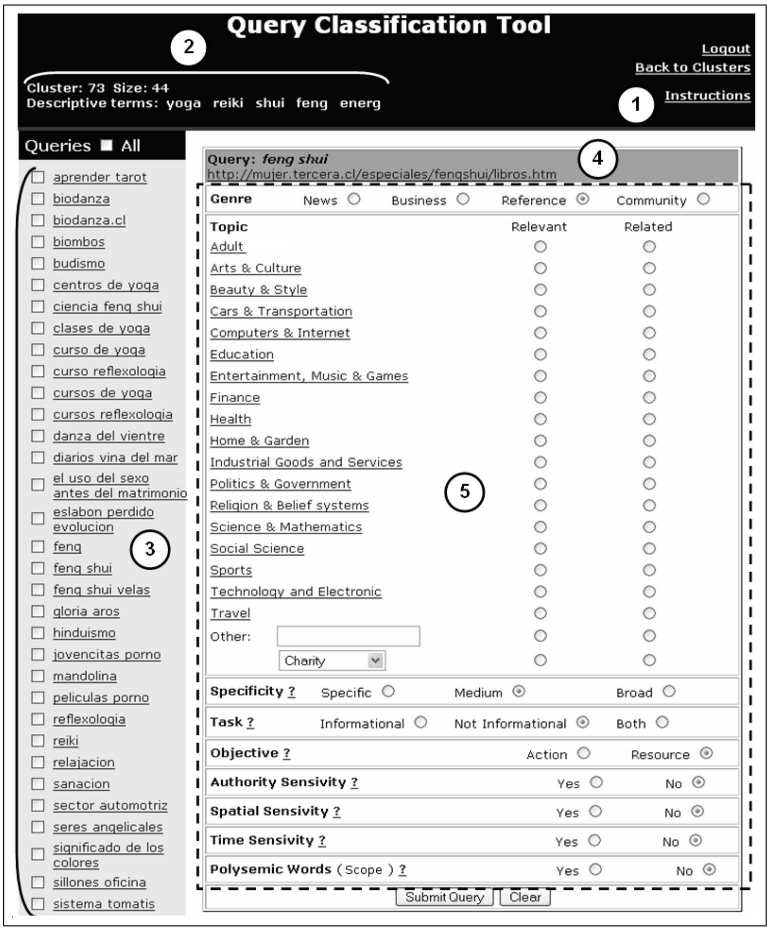
In order to perform the classification of the queries, we developed a software tool considering our experience with the tool developed for 4. The Figure 5.2 is a screen shot of the "Query Classification Tool". After testing different models and layouts, the design of the main screen was divided in five areas, each one containing valuable information to the classifiers. The five areas are:

1. **Instructions.** Contains information about the classification process as well as some illustrative examples.

2. **Information about the cluster.** This area shows the number of queries grouped in a cluster (i.e., the cluster that is being classified), as well as the descriptive terms of such a cluster. With this information the annotator can make a picture of the subject of the queries contained in this group. In Figure 5.2 the cluster 73, contains 44 queries, and its descriptive terms are: *yoga*, *reiki*, *shui*, *feng*, *energ*.

3. **Queries in the cluster.** This area contains the set of queries that belong to the cluster that is being classified. This scheme allowed the annotators to contextualize a query through finding sets of related queries, and to overcome the problem of having to infer the meaning of queries only from their terms.

   There are very compact clusters where most of the queries are very similar. In these situations, the annotator can select such queries and classify all them at the same time. For example, in Figure 5.2 the queries *yoga classes*, *yoga course*, and *reflexology course* can be classified in the same way (i.e., the values for the facets are the same).



Cluster: 73 Size: 44
Descriptive terms: yoga reiki shui feng energ

Queries ■ All

- ☐ aprender tarot
- ☐ biodanza
- ☐ biodanza.cl
- ☐ biombos
- ☐ budismo
- ☐ centros de yoga
- ☐ ciencia feng shui
- ☐ clases de yoga
- ☐ curso de yoga
- ☐ curso reflexologia
- ☐ cursos de yoga
- ☐ cursos reflexologia
- ☐ danza del vientre
- ☐ diarios vina del mar
- ☐ el uso del sexo antes del matrimonio
- ☐ eslabon perdido evolucion
- ☐ feng
- ☐ feng shui
- ☐ feng shui velas
- ☐ gloria aros
- ☐ hinduismo
- ☐ jovencitas porno
- ☐ mandolina
- ☐ peliculas porno
- ☐ reflexologia
- ☐ reiki
- ☐ relajacion
- ☐ sanacion
- ☐ sector automotriz
- ☐ seres angelicales
- ☐ significado de los colores
- ☐ sillones oficina
- ☐ sistema tomatis

Query: *feng shui*
http://mujer.tercera.cl/especiales/fengshui/libros.htm

Genre News ○ Business ○ Reference ● Community ○

Topic	Relevant	Related
Adult	○	○
Arts & Culture	○	○
Beauty & Style	○	○
Cars & Transportation	○	○
Computers & Internet	○	○
Education	○	○
Entertainment, Music & Games	○	○
Finance	○	○
Health	○	○
Home & Garden	○	○
Industrial Goods and Services	○	○
Politics & Government	○	○
Religion & Belief systems	○	○
Science & Mathematics	○	○
Social Science	○	○
Sports	○	○
Technology and Electronic	○	○
Travel	○	○
Other:	○	○
Charity	○	○

Specificity ? Specific ○ Medium ● Broad ○

Task ? Informational ○ Not Informational ● Both ○

Objective ? Action ○ Resource ●

Authority Sensivity ? Yes ○ No ●

Spatial Sensivity ? Yes ○ No ●

Time Sensivity ? Yes ○ No ●

Polysemic Words (Scope) ? Yes ○ No ●

Figure 5.2: Query classification tool.

4. **Query to classify.** In this area, together to the query that is being classified, appears the URL of the most popular Web page that was clicked as answer for the query. In Figure 5.2 the current query is *feng shui*, and its related URL is `http://mujer.tercera.cl/especiales/fengshui/libros.htm`.

5. **Facets.** The facets and their values are contained in this area. If the annotator has a doubt about the meaning or content of a facet, he/she can find related information by clicking on the question mark that appears beside the name of the facet. For example, by clicking on the *topic=Education* a list with the subtopics related to *Education* appears (i.e., Higher Education, Teaching, Homework Help, Preschool, among others).

In spite that using the classification tool is very intuitive, and that it contains basic information about its usage, the annotators were instructed about how to perform the classification of the queries, as well as about the meaning, the content, and the values that each facet could take.

5.5 Analysis of the Query Classification

With the purpose to determine the reliability and consistency of the manual classification, 10% of the queries (i.e., 523 queries) was selected to be classified by two annotators. In order to measure the consistency of the annotators, two well known metrics were applied: the Overall Agreement [48] and the Kappa Coefficient [27, 81]. The former metric expresses a score of how much homogeneity, or consensus, there is in the ratings given by the annotators. The Overall Agreement does not take into account the agreement that would have been expected due solely to chance, so to overcome this problem, we also used the Kappa Coefficient, which is the proportion of agreement between annotators, after chance agreement is removed from consideration.

Albeit both metrics are designed to measure the agreement between annotators, i.e., each of which, independently, classifies each item of a sample into one of a set of mutually exclusive and exhaustive categories, each measure is calculated in a different way, as described below:

Overall Agreement: It is calculated as the number of items on which the annotators agree (which, in this case, corresponds to the queries classified by

Dimension	**Overall Agreement**	**Kappa Coefficient**
Time S.	99.23	0.98
Scope	96.74	0.93
Objective	92.54	0.85
Authority S.	84.32	0.69
Topic	68.26	0.66
Task	75.71	0.63
Spatial S.	81.07	0.62
Genre	65.00	0.53
Specificity	55.44	0.33

Table 5.2: Inter-annotation Agreement and Kappa Coefficient for the dimensions of user's intent.

the annotators in the same manner), divided by the total number of items (that is, the total number of queries).

Kappa Coefficient: In this work the Kappa Coefficient was calculated following the formula of the Free-Marginal Multirater Kappa, taken from [81], as following:

$$\kappa = \frac{\left[\frac{1}{N\ n(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{R} n_{ij}^{2} - N\ n\right) - \frac{1}{R}\right]}{1 - \frac{1}{R}} \qquad (5.5.1)$$

where N is the number of items (i.e., queries classified), n is the number of annotators, and R is the number of rating categories.

Table 5.2 shows the overall agreement of judgements given by the annotators, as well as the Kappa values for each facet or dimension.

5.5.1 Overall Agreement Results

The results from the overall agreement indicate a highly satisfactory consistency of the manual classification. On average, the overall agreement is $\approx 80\%$. Eight out of the nine facets have reached an overall agreement higher than 65%, which is quite high if we consider the number of dimensions that were assessed, the number of possible values that each dimension can take, as well as the different criteria and subjectivity of the annotators.

There are some facets where the agreement among the experts was higher than 90%, this is the case of *time sensitivity*, *scope* and *objective*. The values to be selected for any of these facets can be derived directly from the query terms, which reduce the rate of ambiguity in the assessments, and allows for greater confidence in the manual classification. In particular, the *objective* of a query is *Action* when the query includes a verb, e.g., used car for sale; otherwise it takes the value *Resource*, e.g., geological maps. There is a second group of facets that achieved noteworthy agreement's percentages between 70% and 90%. In this group we find the facets *authority sensitivity, spatial sensitivity, task* and *topic*. It is important to highlight that, despite the fact that the number of possible values that each one of this dimensions can take is high (particularly the *topic* dimension), the level of agreement reached by the experts was high. This is an indicator that queries have some particular characteristics that are related to these facets, and that those characteristics can be identified. Finally, in dimensions such as *genre* and *specificity*, the subjectivity of the expert plays a crucial role. For these dimensions the overall agreement was lower than 70%.

5.5.2 Kappa Coefficient Results

In order to determine the extent to which the observed agreement exceeds what is expected to obtain by chance, we used the Kappa coefficient, that takes values between -1 and 1. There is not a consensus, however, for the interpretation of the strength of agreement for such coefficient. According to the literature, one of the most common used interpretations is the one proposed by Landis and Koch

[64, 86], which suggest that: agreement $\kappa \leq 0$ is a systematic disagreement among the raters; $\kappa = 0$ is random; $0.01 \leq \kappa \leq 0.20$ is slight; $0.21 \leq \kappa \leq 0.40$ is fair; $0.41 \leq \kappa \leq 0.60$ is substantial; $0.61 \leq \kappa \leq 0.80$ is moderate; and $0.81 \leq \kappa \leq 1$ is almost perfect.

In light of the above interpretation, the agreement of the annotators for the facets is: *time sensitivity, scope* and *objective* are almost perfect; the values obtained for *authority sensitivity, topic, task* and *spatial sensitivity* are substantial; the agreement for the facet *genre* was moderate; and the consensus for *specificity* is fair.

The values obtained from Kappa coefficient reflect a consistency and reliability of the manual classification. The assessment from the annotators for all the facets is beyond chance, and even for three facets, the agreement was almost perfect.

The agreement for the last two facets, from Table 5.2, is not very high. This leads us to consider a deeper analysis of the values that each of these facets can take, in terms of their exclusiveness and coverage. As it is usual in this kind of studies, the subjectivity of the annotators plays a special role, and was more evidenced in the classification of the *specificity* dimension than the others. This justifies the slight values obtained for the overall agreement as well as the Kappa coefficient. If this facet is the hardest for people, also explains why is also hard to predict automatically [18].

5.6 Classification Capability of the Facets

With the purpose to relate the different facets, and to discover possible patterns that co–occur in the data set, we have generated several association rules using the Apriori algorithm [102]. Additionally, in order to determine which dimensions are more interconnected, and should be considered in the task to infer others, we have evaluated the studied dimensions through the Correlation-based Feature Selection algorithm (CFS) [43]. From this analysis we have more elements to determine which the most informative facets are. In this work, we used the Weka implementation of both algorithms [102].

5.6.1 Best Facet Predictors

The Correlation-based Feature Selection algorithm (CFS) evaluates the quality of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. The heuristic by which CFS measures the "goodness" of feature subsets takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them.

Hall *et al.* [43] propose the following hypothesis on which the heuristic is based: "Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other". This hypothesis is formalized as follows:

$$G_s = \frac{k\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii'}}}} \tag{5.6.2}$$

where k is the number of features in the subset; $\overline{r_{ci}}$ is the mean feature correlation with the class, and $\overline{r_{ii'}}$ is the average feature intercorrelation.

Through this algorithm, the subsets of features that are highly correlated with the class, and have low inter correlation, are preferred.

In this work, we used the Weka [102] implementation of the CFS algorithm which was configured to use the Greedy Stepwise search method.

By taking the top four better predictive dimensions for each facet, we found that *topic*, *task*, *objective*, and *genre* are, in general, the most informative ones. According to these results, we can predict *genre* by using *topic*, *task*, and *objective*. Facets such as *authority sensitivity*, *scope* and *specificity* provides information that helps to polish the classification. For example, *authority sensitivity* and *scope* are the best predictors for *specificity*. On the other hand, *time sensitivity* is a dimension that the algorithm found not useful as predictor for any other facet. The fact that a very low number of the classified queries took a positive value for this dimension, makes it not discriminative, and, consequently, in case of an automatic classification, this is a good candidate to be removed from the set of facets. On the other hand, time is a completely independent dimension.

5.6.2 Association Rules

Association rules are *if-then rules* that have their origin in market basket analysis. Association rules are now one of the most popular tools in data mining [44].

Association rules have two measures which quantify the support and confidence of the rule for a given data set [102]. These measures are defined as:

- **The coverage** of an association rule is the number of instances for which it predicts correctly. The coverage is often called *support*.
- **The accuracy**, which is often called *confidence*, is the number of instances that a rule predicts correctly, expressed as a proportion of all instances it applies to.

Since the first and arguably most influential algorithm for efficient association rule discovery is Apriori, in this work we selected this model to generate association rules in order to discover different regularities that underlie the set of queries.

The Apriori Algorithm is based on the following key concepts [29]:

- **Frequent Itemsets.** The sets of items which have minimum support (denoted by L_i for th ith-Itemset).
- **Apriori Property.** Any subset of a frequent itemset must be frequent.
- **Join Operation.** To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself.

The Apriori Algorithm is an influential algorithm for mining frequent itemsets for boolean association rules. In order to generate association rules, the Apriori Algorithm follows the steps presented in Algorithm 5.3.

Algorithm 5.3 Apriori Algorithm

```
Step 1: Find the frequent itemsets: the sets of items that have minimum support.
A subset of a frequent itemset must also be a frequent itemset, i.e.,
if $\{AB\}$ is a frequent itemset then
    both $\{A\}$ and $\{B\}$ should be a frequent itemset.
end if
Iteratively find frequent itemsets with cardinality from 1 to $k$ ($k$–itemset).
Step 2: Use the frequent itemsets to generate association rules.
Step 3: Join Step: Candidate itemset of size $k$, $C_k$, is generated by joining $L_{k-1}$ with itself.
Step 4: Prune Step: Any $(k-1)$–itemset that is not frequent cannot be a subset of a frequent $k$–itemset
```

In order to generate rules that consider the largest possible number of dimensions, the algorithm was configured to allow rules with minimum support of 0.01 and confidence higher than 80%. Table 5.3 contains some of the most interesting rules that we found.

There are several rules with very high support, but here we show only some examples of the most outstanding ones. A particular case to outline is the rule number 1 where, even though, its confidence is not 100%, almost 60% of the queries that were labeled as *task=Informational*, have the *objective=Resource*, instead of performing an *Action* (that is the other value that the *objective* dimension can take).

In general terms –and as it was determined through the CFS algorithm–, we can establish that the facets *task* or the *objective* for a user query can be inferred from the dimensions *genre* and *topic*. For example, rules whose *genre* is *Business* and the *topic* is oriented towards finding a product or service (such as *Cars & Transportation* or *Travel*), are related to the *Action objective*; see rules 2–4 and 6. In particular, rule number 2 evidences an intention from the user to

	Lhs		Rhs	Support	Confidence
1.	task=*Informational*	→	objective=Resource	0,590	0,99
2.	genre=Business, topic=Cars & Transportation, specificity=Specific, objective=Action	→	authority sensitivity=Yes	0,017	1
3.	genre=Business, topic=Industrial Goods & Services, task=Not-*Informational*, specificity=Medium	→	objective=Action	0,016	1
4.	genre=Business, task=Ambiguous, specificity=Specific, objective=Action, authority sensitivity=Yes, spatial sensitivity=Yes	→	topic=Cars & Transportation	0,016	0,99
5.	genre=News, topic=Entertainment Music & Games, specificity=Specific, authority sensitivity=No	→	task=*Informational*	0,012	1
6.	genre=Business, topic=Travel, specificity=Medium, objective=Action	→	task=Not-*Informational*	0,012	1
7.	topic=Technology & Electronic, specificity=Medium, spread=No	→	spatial sensitivity=No	0,012	1
8.	topic=Social Science, task=*Informational*, specificity=Medium, authority sensitivity=Yes	→	genre=Reference	0,011	1
9.	genre=Reference, topic=Religion & Belief Systems, objective=Resource, authority sensitivity=Yes	→	task=*Informational*	0,010	1
10.	genre=Reference, topic=Health, objective=Action	→	authority sensitivity=Yes	0,010	1

Table 5.3: Association rules generated from the query labels.

establish contact (or to make a transaction) with a manufacturer or cars dealer (or auto parts) to look for a specific model and a brand he/she has in mind.

Queries where the *topics* are oriented to humanities (such as *Social Sciences* or *Religion & Belief Systems*) appear together with an *author sensitivity* and have a *Medium specificity*. These queries are more intended to obtain information than carrying out an *Action*. Some examples of this can be seen in rules 5 and 7–10.

5.7 Summary

Web search engines tend to view the query formulation and the retrieval process as a simple goal-focused activity. However, the intentions behind Web queries are more complex and the search process should be more oriented to a variety of characteristics than only to a simple goal. In this Chapter we have introduced, analyzed and characterized a wide range of factors or dimensions that may be useful for user's intent identification when searching for information on the Web. These dimensions/facets are: *genre, topic, task, objective, specificity, scope, authority sensitivity, spatial sensitivity*, and *time sensitivity* (Section 5.2).

We have described the main features of each dimension (Section 5.2), their usefulness (Section 5.3), and analyzed some relationships among them (Section 5.6). From the analysis of the manual classification of queries we found that, to certain level, most dimensions plays an important role in the user's intent identification. Dimensions such as *genre, topic*, or *objective* are easier to determine than *task* and *specificity*. To classify a query, the former dimensions require a lower level of subjectivity from the experts, in comparison to the latter ones.

We have confirmed the benefit to separate specific *topics*, from those that have a more general purpose. Dimensions with general purpose may be considered as part of *genre*; hence, we proposed that *genre* is a *meta*–dimension which has a direct impact on the other facets. The use of this meta–dimension allows a fine-grained classification of queries at the *topic* level, as well as a more direct relation with the *objective* pursued by the user. For example, regardless the *topic* of a query, the *Business genre*, in most of the cases, is related to perform an *Action*, which is one of the values of the *objective* dimension.

Additionally, we have analyzed the prediction capability of the facets and determined which ones are the stronger dimensions as well as the weaker ones (Section 5.6.1). From this analysis we found that the former set of facets, i.e. the set of the best predictors is compound by *topic, task, objective* and *genre*. With respect to the facets weakest predictive, the CFS model did not found *time sensitivity* as informative (or discriminative) for the prediction of the others. In general terms, the rules generated (Section 5.6.2) follow a behavior that confirm the findings obtained from the CFS.

Given that the analysis of the multi–dimensional representation of the user intent presented in this Chapter can be considered the basis for automatic identification of the user's intent, in the following Chapter we present a framework towards an on-line identification of the user's intent in Web search.

Chapter 6

Towards On–line Identification

Call to me and I will answer you and tell you
great and unsearchable things you do not know.
Jeremiah 33:3

6.1 Introduction

The problem of identifying user intent behind search queries has received considerable attention in recent years, particularly in the context of improving the search experience via query contextualization [57, 98, 99]. The traditional classification of user intent as *navigational*, *transactional* or *informational* has focused on intent as a single dimension [28] or hierarchical extensions [18, 49, 82]. However, as we said in the last Chapter, the *intent* can be characterized by multiple dimensions, which are often not observed from query words alone. Accurate identification of intent from query words remains a challenging problem primarily because it is extremely difficult to discover these dimensions. The problem is often significantly compounded due to lack of a representative training sample.

There is a need for a systematic framework to analyze user intent in terms of multiple dimensions (or *facets*) of interest. In fact, if a facet can be predicted well, it is possible to contextualize the answer for a query, triggering a different ranking algorithm or even a completely different user experience. For example, an *informational*l query will rank better information coming from trusted Web sites or a time sensitive query may give better ranking to news. Similarly, a local query may trigger a map or a genre specific query may show a particular screen layout. The multiple facets of interest might be correlated, which poses a challenge compared to past approaches [18, 53, 65]. Further, one should be able to predict the user intent for new queries in a reasonably fast manner which can be incorporated into existing search engine technology [34].

In this Chapter we present a generic, extensible framework for learning the multi-dimensional representation of user intent from the query words. The approach models the latent relationships between facets using tree structured distributions for identifying the multi-faceted intent of users based on just the query words. We also incorporated WordNet to extend the system capabilities to queries which contain words that do not appear in the training data. We model the probabilistic dependencies between the various *dimensions* underlying the user intent of a search query, and then we propose to use the model to identify the *intent* behind the query. By modeling the dependencies in an adequate way we should get a more accurate identification of the underlying intent. We use the set of facets from the previous Chapter to illustrate our method. As our model is more complex, we explore the potential of our technique using the minimal set of features available, that is, just the query words. Using more features, such as clicks, would just improve our results.

Our approach immediately raises several challenging questions, namely (a) What kind of models should one use for modeling dependencies? (b) How does one model the joint distribution of query words and the dimensions? and (c) Is it possible to augment the model to cater for unseen query words?

We pose the problem of identifying the user's intent behind a query as an inference problem over a tree structured graphical model. We argue that the tree structured model serves as a natural choice for discovering latent dependencies between several *dimensions* of a query as it is the *simplest* distribution which goes beyond the independence assumption. The associated learning problem is solved by using the Chow-Liu algorithm [31, 70, 97]. In addition, the resulting algorithm should be scalable and preferably on–line for real world use.

Most of the works available in the literature are based on the analysis of query logs and click-through data. However, the additional information might not be available, or restricted due to privacy concerns. Our method infers the user intent from the query words only, without using click-through documents, Web session information, and other meta-data. This shows the potential of the algorithm, as with more features, the performance should improve.

We model the query as a factorial distribution over the observed *query words*, which is equivalent to making the assumption that the impact of a word in the query is independent of the other words. This results in a simplified model which allows an on–line algorithm for user's intent identification. We also incorporate WordNet [77], a lexical database, to improve the classification accuracy of the user intent of search queries for unseen query words.

In order to perform the identification of the underlying user's needs in the described setting, we introduce *FastQ*, an on–line algorithm for classifying user intent for search queries. The algorithm dynamically constructs a tree structured factor graph [63] with the appropriate distribution. The intent is inferred by using standard belief propagation (BP) algorithms, and the tree structure of the dynamically constructed factor graph guarantees the convergence of the BP algorithm. Further, the small size of the resulting factor graph corresponding to a search query allows fast (approximate) inference and allows on–line classification of the user's intent. Experimental timing results show that the method can be incorporated into existing search engine technology [34, 30].

6.1.1 Contributions

The main contributions of this Chapter are:

- We cast the problem of identifying user intent for Web search queries as an inference problem; and present a tree structured graphical model for modeling the user intent based on just the query words.

- We incorporate WordNet [77], a lexical database, to identify new words present in the search query which are absent in the existing training vocabulary.

- Our framework is generic and naturally lends itself to extension by using other facets or query features, and thus enabling its use in other settings.

6.1.2 Outline

The reminder of this Chapter is organized as follows: Section 6.2 presents the dimensions that compound the user's intent. Section 6.3 describes the query intent model for the Web query classification. Section 6.4 presents our experimental setup and the results obtained. Finally, Section 6.6 presents a summary of the Chapter and the most important findings of this work.

6.2 Dimensions of the User's Intent

We chose the multi-dimensional description of user intent presented in Chapter 5. A summary of the used dimensions, as well as the values that each one of them can take, is presented in Table 6.1. The multi-dimensional based description encapsulates additional information about the user intent in a number of related dimensions, instead of a hierarchical scheme.

6.3 Mathematical Model

FastQ casts the problem of identifying the intent of a user query in terms of meaningful facets based on the words in the query as an approximate inference problem. This is complicated by the latent relationships between the facets.

Dimension	**Values & Meaning**
Genre	{*News, Business, Reference, Community*}. It is considered a meta-category that provides a generic context to the user query intent.
Topic	{*Adult, Arts & Culture, Beauty & Style, Cars & Transportation, Computers & Internet, Education, Entertainment, Music & Games, Finance, Food & Drink, Health, Home & Garden, Industrial Goods & Services, Politics & Government, Religion & Belief systems, Science & Mathematics, Social Science, Sports, Technology & Electronic, Travel, Work*}. This list of topics was created by taking the first level categories included in ODP, Yahoo!, and Wikipedia.
Task	{*Informational, Not Informational, Ambiguous*} [18]. Considered as the primary need reflected by a query.
Objective	{*Action, Resource*}. Represents the aim of a query, without considering the format of the information to retrieve.
Specificity	{*Specific, Medium, Broad*}. This facet describes how specialized is a query.
Scope	{*Unique, Multiple*}. Shows whether the query contains polysemic words or not.
Authority Sensitivity	{*Yes, No*}. Is the query designed to retrieve authoritative and trusted answers? [75]
Spatial Sensitivity	{*Yes, No*}. Reflects the interest of the user to obtain a resource that is related to a particular spatial location (explicit or not).
Time Sensitivity	{*Yes, No*}. Whether the information to retrieve, involves a date or period of time.

Table 6.1: The dimensions or *facets* which characterize user intent.

Further, successful integration of query intent requires a fast on–line procedure which can be used as a pre-processing input to improve search engine results.

FastQ models the latent relationships between facets as a second-order product distribution. It incorporates the evidence of the query words on the underlying facet classification under the simplifying assumption that the evidence of each query word on a facet can be modeled independent of the evidence of other words. The resulting factorization results in a factor graph without cycles. This allows exact solution of the underlying inference using standard Belief Propagation (BP) algorithms.

The model constructs a tree factorization of the latent facet interrelationships based on the Chow-Liu algorithm and using a small number of manually classified training examples by multiple experts with high degree of confidence. The Chow-Liu algorithm generates the maximum likelihood estimate (MLE) of possible second-order product distributions based on the empirical distribution observed in the training data.

Modeling the impact of query words on the facet classification presents on additional challenge due to the small number of available training samples. *FastQ* uses a Bayesian approach by using appropriate priors and using maximum a posteriori (MAP) estimate to obtain the factorization which models the impact of a query word on the facet classification independent of other query words. Further, the results are augmented to handle words not present in the existing vocabulary using the semantic relationships between words by incorporating the WordNet [77] lexical database to augment evidence present by the new word in terms of its semantic neighbors.

6.3.1 User Intent Classification

This Section presents a fast on–line procedure for the user intent classification of a test query as:

$$\hat{\vec{f}}^{\dagger} = \arg\max_{\vec{f} \in \mathcal{F}} P(\breve{Q} = \breve{q}^{\dagger} | \vec{F} = \vec{f}, \hat{\Lambda}^{*}) P(\vec{F} = \vec{f} | \Theta_T^{*}) \tag{6.3.1}$$

A naive solution requires the computation of $O(\prod_{i \in \mathcal{K}} \mathcal{F}_i)$ product terms, each having $O(L|\mathcal{K}|)$ multiplications. Further, the computation of $Z(\vec{f})$ for each $\vec{f}$ requires a summation over $\mathcal{V}^m$ terms, which is computationally infeasible.

Instead, we perform approximate inference using standard Belief Propagation (BP) on a dynamically constructed factor graph. Figure 6.1 shows a possible

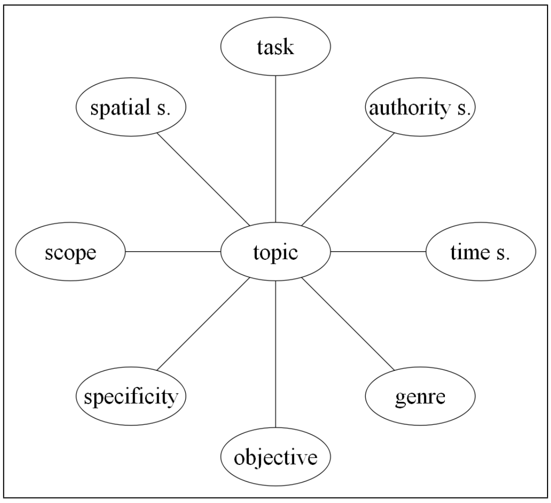


Figure 6.1: Example spanning tree (T) on facets.

tree structure T defined on the facet variables. The resulting factor graph is tree structured having K variable nodes and $O(MK)$ factor nodes, and the inference procedure converges in at most K iterations.

Dynamic Factor Graph

The factor graph for finding the optimal facet assignment $\vec{f}^{\dagger}$ for a test query $\breve{q}^{\dagger}$ consists of two classes of factors, namely,

- factors which model the latent relationship between facets.
- factors which model the impact of query words on facets, under the simplifying assumption that there are no latent relationships between facets.

Figure 6.2 shows the factor graph corresponding to the facet classification. The variable nodes, $F_1, \ldots, F_K$, denote the facet variables which can take possible states in $\mathcal{F}_1, \ldots, \mathcal{F}_K$, respectively. The factors T_i, T_{ij} model the latent relationship between facets (corresponding to the Chow-Liu probability distribution on facet variables.), while the factors W_ℓ^i model the impact of the ℓ^{th} word of the test query $\breve{q}^\dagger$ affecting the i^{th} facet variable.

The factors, $T^i, i \in \mathcal{K}$ and $T^{ij}, (i,j) \in T^*$ correspond to the latent relationships between the facets. Their beliefs correspond to the parameters (Θ^*, T^*) based on the training data, given as

$$\begin{aligned} P_{T^i}(F_i = f) &= \Theta_i^*(f)^{(1-d_i)} && (6.3.2) \\ P_{T^{ij}}(F_i = f, F_j = f') &= \Theta_{ij}^*(f, f') && (6.3.3) \end{aligned}$$

The factors W_ℓ^i model the conditional distribution $\hat{P}(F_k = f | w \in \breve{q})$ under the assumption that all facets are independent of each other. The factors are specified by the beliefs,

$$P_{W_\ell^i}(F_i = f) = \xi_{w;f}^i = \frac{\#(\vec{f_k} = f, w \in \breve{q})}{\#(w \in \breve{q})} \tag{6.3.4}$$

where $\xi_{w;f}^k$ is an approximation for $\hat{P}(F_k = f | w \in \breve{q})$ obtained as

$$\xi_{w;f}^k \quad \propto \quad P(w \in \breve{q} | F_k = f, \hat{\Lambda}^*) \hat{P}_1^k(F_k = f) \tag{6.3.5}$$

and $\hat{P}_1^k(F_k = f) = \frac{\#(\vec{f_k}=f)}{N}$ is the MLE estimate of the probability distribution of the k^{th} facet under the simplifying assumption that the each facet is independent of other facets.

The parameter $\Xi^* = \{\xi_{w;f}^k : f \in \mathcal{F}_k, k \in \mathcal{K}, w \in \mathcal{V}\}$ models the impact of the words in a query on the facets. Specifically, $\xi_{w;f}^k$ measures the evidence that the k^{th} facet has value f when the query contains the word w.

We note that the evidence $\xi_{w;f}^k$ in (6.3.4) is not defined if the word w is not part of the training vocabulary, i.e., the test query $\breve{q}^\dagger$ contains a word w that is not present in the training data $\mathcal{D}$. The following Section presents a method for incorporating the influence of new words, i.e. words in the test query which are not present in the vocabulary using the semantic relationships between the words and the existing vocabulary.

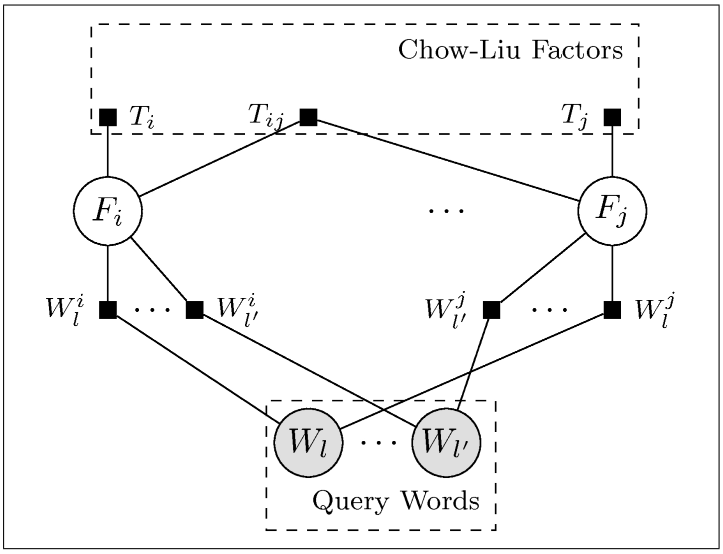


Figure 6.2: Factor graph for query facet classification.

WordNet Integration

Let w denote a new word not present in the vocabulary $\mathcal{V}_{\mathcal{D}}$. Then, the corresponding $\xi^i_{w;f}$ are not defined and the previous algorithm cannot be used. However, a semantically related word w' might be present in the vocabulary $\mathcal{V}_{\mathcal{D}}$; for which the corresponding $\xi^i_{w';f}$ are defined. It is safe to assume that a strong semantic relationship between the words w and w' will be reflected in $\xi^i_{w;f}$ and $\xi^i_{w';f}$. Therefore, it is of interest to be able to relate new words in the query to existing words in vocabulary.

Now we present a two-step method for integrating the capabilities of WordNet [77], a lexical database, into the facet classification for queries containing new words. We first query WordNet for the related word candidates $W^{\perp}$ with similarity scores $S^{\perp} = \{s_{w'} : w' \in W^{\perp}\}$ for a new query word w using Breadth First Search (BFS) till a maximum search depth L. The second step uses the candidates which are present in the vocabulary $W^{\clubsuit} = W^{\perp} \cap \mathcal{V}_{\mathcal{D}}$ to compute the parameter $\tilde{\xi}^i_{w;f}$. Algorithm 6.4 presents the heuristic procedure for incorporating

Algorithm 6.4 WordNet integration $\tilde{\xi}_w = F_{\mathrm{WN}}(w)$.

```
Require: $L$ {Maximum depth for BFS}
Require: $\Xi^*$ {Parameter based on $\mathcal{D}$}
Ensure: $w \notin \mathcal{V}_{\mathcal{D}}$ {New word not present in vocabulary}
Ensure: $\tau_f^i = \frac{\#(f_i=f)}{N}$
  $(W^\perp, S^\perp) \leftarrow BFS(w, L)$ {Related words in WordNet}
  $W^\clubsuit \leftarrow W^\perp \cap \mathcal{V}_{\mathcal{D}}$ {Related words in vocabulary}
  for $i = 0$ to $K$ do
    Initialize $\{\mu_1^i, \ldots, \mu_{|\mathcal{F}_i|}^i\} \sim Dir(\tau_{f_1}^i, \ldots, \tau_{f_{|F_i|}}^i)$
    for all $f$ in $\mathcal{F}_i$ do
      $\tilde{\xi}_{w;f}^i = \frac{\mu_f^i + \sum_{w' \in W^\clubsuit} s_{w'} \xi_{w';f}^i}{\sum_{f \in \mathcal{F}_i} (\mu_f^i + \sum_{w' \in W^\clubsuit} s_{w'} \xi_{w';f}^i)}$
    end for
  end for
  return $\tilde{\xi}_w = \{\tilde{\xi}_{w;f}^i : i \in \mathcal{K}, f \in \mathcal{F}_i\}$
```

related words obtained from WordNet to augment results for a new word.

We note that if the word has no neighbors either in WordNet or in the existing vocabulary, the conditional distribution $P(F_i = \cdot | w \in \breve{q}^\dagger)$ is a Dirichlet distribution with parameter:

$$\left[\frac{\#(F_i = f_1)}{N}, \ldots, \frac{\#(F_i = f_{|\mathcal{F}_i|})}{N}\right] \tag{6.3.6}$$

This means that when the test word w has no semantically related neighbors w' with known statistics $\xi_{w';f}^k$, then the conditional distribution is randomly distributed such that, on average, the count statistics for a facet would be same as that found in the training data. However, if there exist neighbors of the word $W^\clubsuit$, that information is incorporated into the conditional distribution for the new word.

Algorithm 6.5 FastQ $\hat{f}^{\dagger} = F(\breve{q}^{\dagger}|\mathcal{D})$.

```
Require: $\breve{q}^{\dagger}$ {Test query}
Require: $\Xi^*, \Theta^*, T^*, \mathcal{V}$ {Parameter from training data $\mathcal{D}$}
  Initialize $G(\mathcal{K}, [])$ {factor graph on facet variables}
  for all $(i, j)$ in $T^*$ do
    Add factor $T^{ij}$ s.t. $P_{T^{ij}}(F_i = a, F_j = b) = \Theta^*_{ij}(a, b)$
  end for
  for all $i$ in $\mathcal{K}$ do
    Add factor $T^i$ s.t. $P_{T^i}(F_i = a) = \frac{1}{\Theta^*_i(a)^{d_a - 1}}$
  end for
  for $l = 1$ to $L$ do
    if $q^{\dagger}_{\ell} \in \mathcal{V}$ then
      for all $i$ in $\mathcal{K}$ do
        Add factor $W^i_{\ell}$ s.t. $P_{W^i_{\ell}}(F_i = f) = \xi^i_{q^{\dagger}_{\ell};f} \forall f \in \mathcal{F}_i$
      end for
    else
      $\tilde{\xi}_{q^{\dagger}_{\ell}} = F_{WN}(q^{\dagger}_{\ell})$ {Use WordNet for new query word}
      for all $i$ in $\mathcal{K}$ do
        Add factor $W^i_{\ell}$ s.t. $P_{W^i_{\ell}}(F_i = f) = \tilde{\xi}^i_{q^{\dagger}_{\ell};f} \forall f \in \mathcal{F}_i$
      end for
    end if
  end for
  $\hat{f}^{\dagger}$ = MAX-PROD($G$) {Standard BP}
  return $\hat{f}^{\dagger} = \{f^{\dagger}_1, \ldots, f^{\dagger}_K\}$
```

The *FastQ* Algorithm

The overall procedure for facet classification when a new query, $\breve{q}^{\dagger} = \{q_1, \ldots, q_L\}$ is given in Algorithm 6.5. The algorithm constructs the dynamic factor graph by adding the factors T_i, T_{ij} corresponding to the probability distribution on facets constructed using the Chow-Liu algorithm. Then, if a query word q_ℓ is in the vocabulary $\mathcal{V}$, it adds words to the corresponding word factors, and constructs the word factor W_ℓ^i based on parameter Ξ^*, as in (6.3.4). Otherwise, if query word q_ℓ is not in vocabulary $\mathcal{V}$, it constructs the factors W_ℓ^i based on semantically related words in vocabulary $\mathcal{V}$ found using WordNet. The facet assignment $\hat{f}^\dagger$ for the test query is computed using a standard implementation of the Max-Product Belief Propagation [63].

The small size of the constructed factor graph allows an implementation with limited resources. The maximum search depth L in WordNet is a user-controlled parameter which can be reduced to speedup the classification.

6.4 Experimental Setting

This Section describes the dataset used for the experiments (Section 6.4.1), and the experimental setup (Section 6.4.2).

6.4.1 Dataset

The evaluation data used in this work are the same used in the Section 5.4. This data consist of $5,249$ queries from a vertical search engine query log in a Spanish speaking country. The set of queries were manually classified in the set of facets that were presented in Section 6.2.

The details about the manual classification, as well as the results of the metrics applied to measure the inter–annotation reliability are described in Section 5.4.

Training and testing sets. The amount of labeled data available in most IR scenarios is much smaller than the corresponding test sets. In order to test the models in the low data regime, we have used 1, 5, 10, 50 $\%$ of the available queries as training data. We have repeated the experiment for $N = 10$ independent trials where the training samples have been randomly chosen. We report the resulting mean and variance across the trials.

	1%			5%			10%			50%		
Model	BASE	WN	FASTQ	BASE	WN	FASTQ	BASE	WN	FASTQ	BASE	WN	FASTQ
Author S.	33.48	43.44	**79.85**	42.35	55.41	**79.76**	50.07	61.64	**81.14**	55.97	64.82	**81.70**
Genre	30.73	35.51	**51.95**	38.53	43.85	**59.76**	43.18	47.42	**60.63**	50.00	52.47	**65.02**
Objective	**81.96**	80.03	73.09	**84.99**	82.51	83.35	**86.03**	83.24	83.01	**86.19**	85.62	85.59
Scope	23.51	45.42	**97.56**	37.88	60.75	**97.51**	46.69	67.67	**97.40**	56.51	72.79	**97.28**
Spatial S.	45.42	45.15	**59.42**	49.01	49.68	**61.62**	52.82	54.39	**62.14**	56.27	58.32	**67.98**
Specificity	75.62	73.55	**77.51**	75.93	74.37	**76.98**	77.24	73.96	**77.51**	77.08	76.41	**78.14**
Task	**68.47**	66.61	61.45	**72.60**	69.16	71.02	**75.21**	71.90	71.64	**76.87**	75.78	76.04
Time S.	21.25	42.54	**97.98**	35.55	59.32	**98.60**	44.61	64.87	**98.64**	54.37	71.38	**98.43**
Topic	11.73	12.85	**14.01**	**21.97**	20.75	19.39	**27.57**	26.90	25.67	34.44	**36.00**	32.90

Table 6.2: Classification accuracy (in %) of the user intent dimensions for the different models.

6.4.2 Models Tested

The two key components of *FastQ* are the modeling of the latent relationships between facets and the incorporation of WordNet lexical database. In order to highlight the impact of the above two factors in the overall results and establish a baseline, we construct two variants of the *FastQ* algorithm described below:

- *BASE*: The model ignores the latent relationships between facets and does not incorporate WordNet to compute the influence of new words in the test query. Thus, the resulting factor graph does not have the factors T_i and T_{ij}; and only has factors W_ℓ^i for the words which are already present in the vocabulary $\mathcal{V}$; ignoring the test query words not present in the vocabulary.

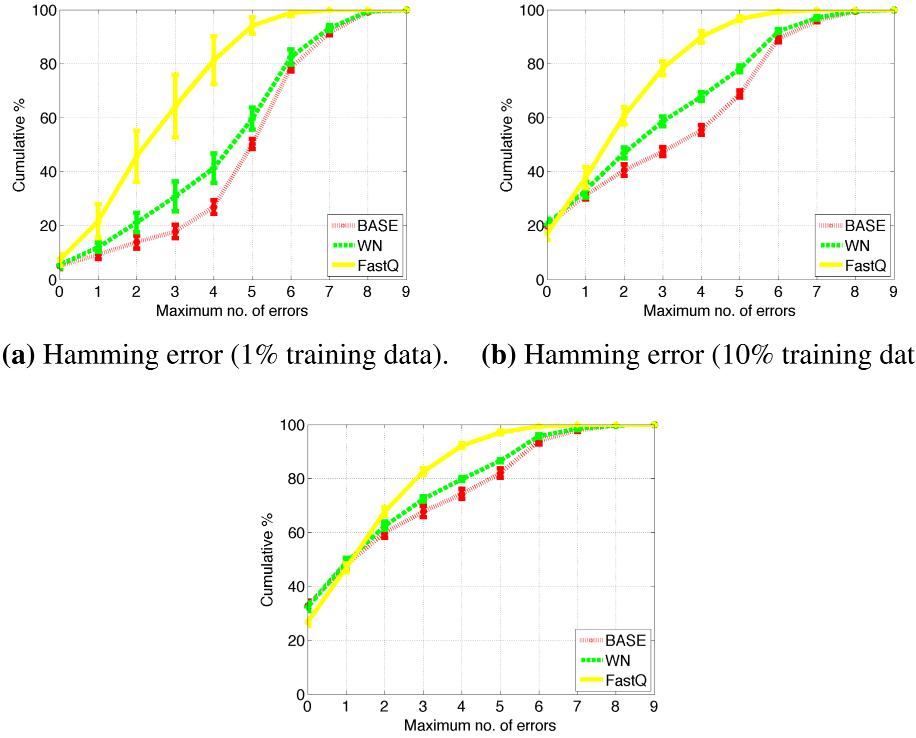
- *WN*: This does not model the latent relationships between facets; but does incorporate WordNet to incorporate the influence of new words in the test query in terms of their semantic neighbors in the existing vocabulary. Thus, the resulting factor graph does not have the factors T_i and T_{ij}; but, has factors, W_ℓ^i for the new words in the test query in addition to factors for known words.

6.5 Results

The evaluation of the models was done addressing three aspects of concern, namely, accuracy of prediction, the impact of WordNet, and the time taken for query classification.

Classification Accuracy

Table 6.2 presents the classification accuracy of the individual facets. We note that the latent modeling of the facets in *FastQ* improves the accuracy compared to *BASE* and *WN* for the facets: Time Sensitivity, Scope, Spatial Sensitivity, Authority Sensitivity, and Genre.



(a) Hamming error (1% training data).

(b) Hamming error (10% training data).

(c) Hamming error (50% training data).

Figure 6.3: The mean (curve) and variance (shown as a vertical bar) of the Hamming error for the overall facet classification for the models *BASE*, *WN*, and *FastQ*.

We use the Hamming error to measure the overall accuracy of the prediction, i.e., for a query $\breve{q}^\dagger$ with true facet classification $\vec{f}^\dagger$ and the algorithm prediction $\hat{f}$; the Hamming error $d_H(\vec{f}^\dagger, \hat{f})$ is $d_H(\vec{f}^\dagger, \hat{f}) = \sum_{k=1}^{K} 1_{\hat{f}_k \neq \vec{f}_k^\dagger}$ where $1_{\hat{f}_k \neq \vec{f}_k^\dagger}$ is 1 if the k^{th} facets are not same; else 0. A low (ideally $zero$) Hamming distance means few (ideally $none$) of the K facets have been incorrectly predicted.

Figure 6.3*(a)*–6.3*(c)* shows the cumulative distribution of the Hamming er-

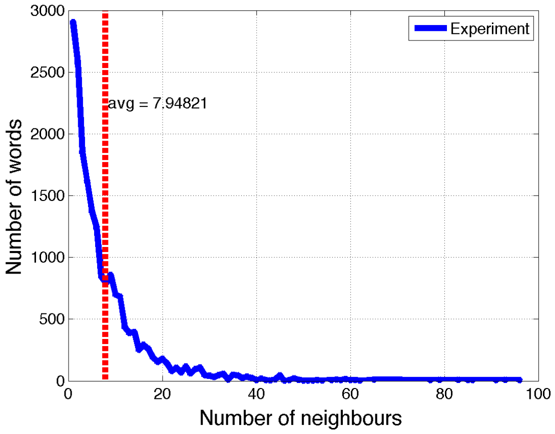


Figure 6.4: Number of related words obtained using WordNet for a test query word, and the expected (average) number of related words for a new query word.

rors (number of individually incorrect facets) in the prediction of K facets for 1%, 10% and 50% training data. For example, when 10% of the data is used for training, *FastQ* classifies over 60% of the test queries with at most 2 mis–classified facets; compared to *WN* (45%) and *BASE* (40%). The modeling of the facet inter-relationships (*FastQ*) improves the overall quality of classification compared to *BASE* and *WN*, especially at low amounts of training data.

We note that FastQ performance compared to BASE is slightly inferior on the facet "Topic", which has 23 possible states. Therefore, accurately modelling the probability distribution requires more data. We note that at 90% training data, FastQ does perform better on "Topic" (and all other facets) compared to BASE. We note that in a real world setting, the amount of available data might be greater than a few thousand queries (for example, when using ODP categories). Under such scenario, FastQ will perform better modeling compared to the results shown here.

We attempt to learn a multi-faceted classification of user intent, which is

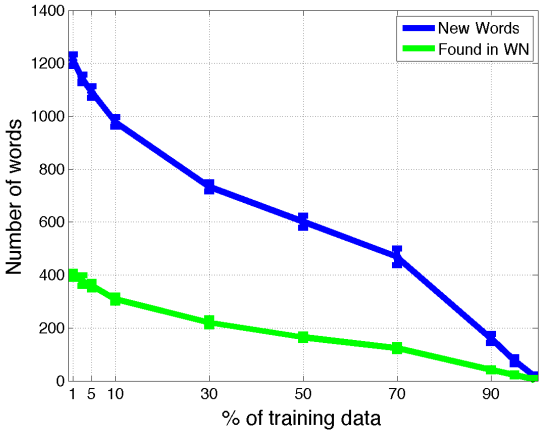


Figure 6.5: Mean (curve) and variance (shown as a vertical bar in each data point) of the number of new words found in the test data, and the number of words for which related words were found in the training data using WordNet with increasing amount of training data.

inherently more sophisticated compared to learning the classical notion of user intent as *informational*, *navigational* or *transactional*. However, the closest analogue in our case is the *Task* dimensions whose values are *Informational*, *Not informational* or *Ambiguous*.

Table 6.3 shows the classification results, in terms of precision (P), recall (R) and the F-measure (F), for the *Task* dimension when 50% of the queries was used as training data, and the remaining for testing. An average accuracy of 76% is achieved by *FastQ* in the classification of *Task*.

Table 6.4 shows that *FastQ* compares favorably to existing methods. The Results of [59] are based on the TREC test collection (‡ in Table 6.4). We reiterate the multi-dimensional nature of our work, i.e. *Task* is one of the nine facets which characterize user intent in our case, compared to past approaches.

	True	**Predicted**	**Precision**	**Recall**	**F**
Informational	1382	1584	0.79	0.91	0.85
Not-Informational	645	635	0.73	0.72	0.72
Ambiguous	212	20	0.89	0.08	0.15

Table 6.3: *FastQ* Classification results for the facet *Task* as *Informational, Not-Informational*, and *Ambiguous.*

6.5.1 Impact of WordNet

A comparison of the results in Table 6.2 and Figure 6.3*(a)*–6.3*(c)* show that the incorporation of WordNet for computing the influence of new words in the test query improves the quality of results in *WN* compared to *BASE*, especially at low amounts of training data ($1 - 10\%$).

Figure 6.5 shows that around 33% of the new words are semantically related to known words in the training data. This might be surprising, but this is explained from the characteristics of the training set that only includes 15% of queries from the head of the query distribution. Thus, incorporating WordNet may allow a better modeling for a significant portion of unseen words in a practical Web search scenario.

Figure 6.4 shows the histogram of the number of related words in WordNet, experimentally observed, for a query word that is not present in training data, and its average $\bar{c}_{wn} \simeq 8$. This affects the time spent to search for related candidates of a word not present in the training data.

The experiments reported used a breadth first search till maximum depth 3 using the words falling in the $synsets$ category. We observed that the similarity scores for the neighbors often is either very close to one, or slightly above zero.

Method	**Size**	**Accuracy**
Topic/homepage [59]‡	200	91%
FastQ (*Task*)	5,249	76%
Rule-based hierarchy [53]	1,500,000	74%
SVM [18]	6,042	63%
Query-log based [65]	50	54%

Table 6.4: Comparison of existing approaches for intent classification with size of the dataset and the classification accuracy.

6.5.2 Efficiency and Scalability

The average time for query classification at WordNet maximum search depth 3 was observed to be $2 - 3$ ms on an AMD Turion64 machine with 4 GB RAM, which is considerably less than the average latency for a search query which is 200 ms [34]. This allows the incorporation of *FastQ* as input to the ranking algorithm of a search engine; using the user intent information to improve the user experience.

In addition, our results show that with a small training set (a few thousand queries), we can get good predictive performance. Hence, with training sets available for large search engines the results might be quite better.

6.6 Summary

We have presented an efficient and convergent algorithm for inferring the intent of the Web search query, described in terms of multiple facets. Experimental results show that modeling the latent relationships between facets improves the accuracy of prediction. Further, the incorporation of WordNet improves intent classification using the semantic relationships in the language vocabulary.

One possible area of improvement is the usage of multiple semantic relationships, e.g. *synonyms, antonyms, hyponyms*, available in WordNet, which might improve the efficacy of WordNet in intent classification. We believe there is scope to further improve the quality of classification by allowing powerful capabilities of WordNet to be leveraged for query intent classification.

Our framework model can also be used for fast contextualized ranking of the results of a query with or without clickthrough data. For example, it can be used for a much more refined query classification and clustering. Similarly, the intent classification can be used as input to existing search engine ranking technology [30]. Alternatively, we can use the facet classifications for sorting and displaying the search engine results in a different way creating a new user experience.

In addition, the generality of this framework allows integration of alternative aspects of user intent making it a powerful tool for contextualizing and personalizing search.

Chapter 7

Conclusions

To every thing there is a season,
and a time to every purpose under the heaven...
Ecclesiastes 3:1

In this Ph.D. work we have focused on identifying and understanding the *intentions* that motivate a user to perform a search on the Web. To this end, we apply machine learning models that do not require more information than the one provided by the very needs of the users, which in this work are represented by their *queries*. The knowledge and interpretation of this invaluable information, can help search engines to obtain resources especially relevant to users, and thus improve their satisfaction.

The different works presented in this thesis, show that is not only possible to identify the user's intentions, but that this process can be conducted automatically by means of *unsupervised learning* techniques. The applied techniques have been selected according to the context of the problem being solved.

Without supervision, the applied techniques have shown to be effective for the problem of identifying user's query intention.

7.1 Results

This thesis has produced the following results:

- In Chapter 3, we have presented the analysis of a Web search engine query log from two different perspectives: the query session and the clicked document. In the first perspective, that of the query session, we processed and analyzed Web search engine query and click data for the query session (query plus clicked results) conducted by the user. In the second perspective, that of the clicked document, we defined possible document categories and selected descriptive variables to define the documents.

- In Chapter 4, we presented a first step for identifying user's intents from a Web search engine's query log. We have proposed that the user's intent can be: *Informational, Not–Informational* and *Ambiguous*. Along to this characterization, in this work we have analyzed a set of topical categories, in which user's intent may be classified. We also made a comprehensive analysis of a set of queries that were manually classified into the user's intent and topical categories, and related these two aspects in order to better identify the motivation of the users in a searching for information process.

- In Chapter 5 we have introduced, analyzed and characterized a wide range of facets or dimensions that may be useful for user's intent identification when searching for information on the Web. These dimensions/facets are: *genre, topic, task, objective, specificity, scope, authority sensitivity, spatial sensitivity*, and *time sensitivity*. We have described the main features of each dimension, their usefulness, and analyzed some relationships among them.

- In Chapter 6, we have presented a generic, extensible framework for inferring the intent of the Web search queries, described in terms of multiple facets. We presented a tree structured graphical model for modeling the user intent based based just on the query words. Through the use of WordNet [77], we have identified new words present in the search query that are absent in the existing training vocabulary.

The presented research is supported by the use unsupervised learning models, whose results have been analyzed with the main purpose to better understand the users, their needs, and the motivations that they lead when search for information on the Web.

7.2 Comparison of Techniques

With the purpose to determine the user's intent in Web search, throughout this thesis we have presented the use of different *unsupervised* learning models. In this section we include a comparison of the behavior of PLSA, that is one of the algorithms used in this thesis (used in Chapter 4), against two algorithms that require additional information (or labeled data), in order to generate predictions or classifications.

- **Supervised learning**. This is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal).

  A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete, see classification) or a regression function (if the output is continuous, see regression). The inferred function should predict the correct output value for any valid input object. For the purpose of this comparison we use the Support Vector Machines algorithm [100].

- **Semi–supervised learning**. This is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically

Class	Supervised			Semi–supervised			Unsupervised		
	Prec.	Recall	F	Prec.	Recall	F	Prec.	Recall	F
Informational	0.646	0.985	0.780	**0.789**	0.771	0.785	0.626	0.723	0.671
Not Inf.	**0.843**	0.219	0.348	0.556	0.605	0.579	0.346	0.421	0.380
Ambiguous	**0.478**	0.021	0.040	0.438	0.423	0.461	0.476	0.277	0.350

Table 7.1: Unsupervised, Semi–supervised and Supervised models for user's query intent prediction.

a small amount of labeled data with a large amount of unlabeled data. For the purpose of this comparison we used the Semi–supervised Linear Support Vector Machines algorithm [87].

The purpose of this comparison is point out some of the strengths and weaknesses of unsupervised learning for the task of determine the user's intent.

Table 7.1 presents the Precision, Recall, and F–Measure of query user intent for 50% of labeled data for Supervised and Semi–Supervised Learning, and 100% of unlabeled data for Unsupervised Learning. The data for this comparison are part of a comparative study of machine learning techniques applied to query intent prediction [17].

Although we use 100% of the query data for the unsupervised learning model, it is important to recall that in this process there is no more information than the user's queries. On the opposite, the supervised and semi–supervised models use 50% of labeled data in order to create a model to predict the intent of the user.

In terms of *Precision*, even though the results of the unsupervised algorithm are not the best, without any supervision, PLSA algorithm is able to identify the same amount of queries that SVM has classified as having an *Informational* and that both, SVM and semi–supervised SVM have classified as *Ambiguous*.

In terms of *Recall*, the unsupervised algorithm outperform the supervised classification for *Not Informational* and *Ambiguous* query intents. Additionally, the amount for *Informational* queries found are quite similar to the semi–

supervised algorithm.

The harmonic mean of Precision and Recall (i.e. *F–measure*) from unsupervised learning model is better than from the supervised one for *Not Informational* and *Ambiguous* query intents.

The values for the semi–supervised learning algorithm, in general, are better than the values for the other two models. This shows that a combination of both supervised and unsupervised models improve the quality of classification. Comparing the supervised and the semi–supervised results, we see how the use of unsupervised data helps in the building of an accurate model.

On the other hand, as we have presented in Section 4.5 of the Chapter 4, the unsupervised learning model could not to determine the complete set of informational categories (i.e., 18 topics). In contrast, the model has found two new, and well represented, informational categories. This is a strength of the unsupervised learning models. Since there is not supervision, this kind of machine learning models works to find natural associations among the data, hence, to discover unseen patterns.

7.3 Future Work

Although in this thesis we have presented several ways of determining the user's intent, this is still an open problem in IR. Figure 7.1 presents several paths that, from our point of view, may be addressed for the automatic identification of user's query intent.

The *representation* of the user is a crucial path to investigate. In this thesis we have established a base to represent the query intent as a *multi–dimensional* problem. However, we consider that it might be enriched by adding other *features* from the user's sessions, and click through data. Also the use of semantic could give more insights to represent the user's query intent.

In Chapter 6 we have proposed a framework for the automatic identification of the user's intent in an *on–line environment*. With such framework we are open-

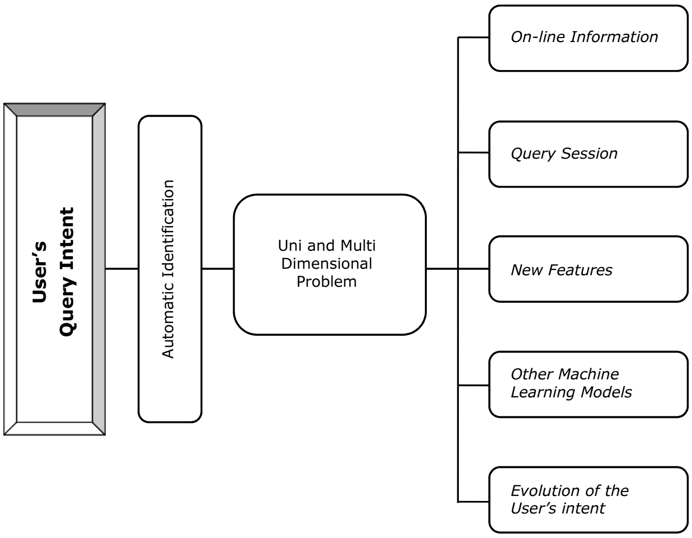

User's
Query Intent
Automatic Identification
Uni and Multi Dimensional Problem
On-line Information
Query Session
New Features
Other Machine Learning Models
Evolution of the User's intent

Figure 7.1: Paths for future work.

ing a door for new works in this direction, whose results can be used by search engines to provide personalization, hence to improve the *user's satisfaction*.

From the last proposal of future work, emerge the need to continue exploring and applying new *machine learning* models to automatically determine the user's intent. With reliable models, the search engines can aid the search process for example by adapting different ranking functions to give more accurate document positioning, adapt the number of answers and user interfaces.

With the continued evolution of the Web, and the creation of social networks such as Facebook [3] and Twiter [8], new data are emerging. The study of this kind data offers a new perspective to analyze the user's intent.

Last but not least, it is important to analyze the *evolution* of the user's intent along the time. This information may help to visualize emerging trends of users needs.

The potential application of user's intent identification is very wide hence, there are many other possibilities of future work in this area, specially for e–commerce, however this is just a sample. ★

Bibliography

[1] Altavista. www.altavista.com.

[2] Excite. www.excite.com.

[3] Facebook. www.facebook.com.

[4] Fireball. www.fireball.de.

[5] Medline. www.ncbi.nlm.nih.gov/pubmed.

[6] Open directory project. dmoz.org.

[7] TodoCL. www.todocl.cl.

[8] Twitter. www.twitter.com.

[9] Wikipedia. en.wikipedia.org.

[10] Yahoo! www.yahoo.com.

[11] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Characterizing query intent from sponsored search clickthrough data. In *Proc. of Information Retrieval for Advertising Workshop at 31th SIGIR Conference*, pages 15 – 22, 2008.

[12] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. In *Proc. of the 31th European Conference on*

Information Retrieval, ECIR, LNCS, pages 578–586, Berlin, Heidelberg, 2009. Springer.

[13] S. Äyrämo and T. Kärkkäinen. Introduction to partitioning–based clustering methods with a robust example. Technical Report C.1/2006, Department of Mathematical Information Technology, University of Jyväskylä, Finland, 2006.

[14] R. Baeza-Yates. Information retrieval in the web: beyond current search engines. *International Journal of Approximate Reasoning*, 34(2-3):97 – 104, 2003. Soft Computing Applications to Intelligent Information Retrieval on the Internet.

[15] R. Baeza-Yates. *Web Mining: Applications and Techniques*, chapter XIV Web Usage Mining in Search Engines, pages 307 – 321. Idea Group, 2004.

[16] R. Baeza-Yates. Applications of web query mining. In *Proc. of the 27th European Conference on Information Retrieval, ECIR*, LNCS, pages 7–22, Berlin, Heidelberg, 2005. Springer.

[17] R. Baeza-Yates, L. Calderón-Benavides, D. Dubhashi, C. González-Caro, and L. Tansini. A comparative study of machine learning techniques applied to query intent prediction.

[18] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In *Proc. of the 13th Symposium on String Processing and Information Retrieval, SPIRE*, volume 4209 of *LNCS*, pages 98–109. Springer, 2006.

[19] R. Baeza-Yates and C. Castillo. Relating web structure and user search behavior (extended poster). In *Proc. of the 10th International Conference on WWW*, Hong Kong, China, 2001.

[20] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology - EDBT*, pages 588–596. Springer-Verlag GmbH, 2004.

[21] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *Proc. of 3rd Latin American Web Congress LA-WEB*, page 242. IEEE Computer Society, 2005.

[22] R. Baeza-Yates and P. Raghavan. Chapter 2: Next generation web search. In S. Ceri and M. Brambilla, editors, *Search Computing*, volume 5950 of *Lecture Notes in Computer Science*, pages 11–23. Springer Berlin / Heidelberg, 2010.

[23] N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

[24] N. J. Belkin. Interaction with texts: Information retrieval as information-seeking behavior. *Information Retrieval'93, Von der Modellierung zur Anwendung*, 93:55–66, 1993.

[25] N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.

[26] N. J. Belkin, R. N. Oddy, and H. M. Brooks. ASK for Information Retrieval. Part I: Background and Theory. *Journal of Documentation*, 38(2):61 – 71, 1982.

[27] R. L. Brennan and D. J. Prediger. Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41:687–699, 1981.

[28] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[29] A. Ceglar and J. F. Roddick. Association mining. *ACM Computing Surveys*, 38, July 2006.

[30] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. *Journal of Information Retrieval*, 13:201–215, June 2010.

[31] C. I. Chow, S. Member, and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

[32] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan. *Data Mining, a knowledge discovery approach.* Springer, New York, NY, 2007.

[33] P. Dayan. Unsupervised learning. In *The MIT Encyclopedia of the Cognitive Sciences*. The MIT Press, 1999.

[34] J. Dean. Challenges in building large-scale information retrieval systems. ACM International Conference on Web Search and Data Mining, WSDM, keynote, 2009.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[36] D. Downey, S. T. Dumais, and E. Horvitz. Heads and tails: studies of web search with common and rare queries. In *Proc. of the 30th International ACM SIGIR Conference*, pages 847–848, New York, NY, USA, 2007. ACM.

[37] D. Downey, S. T. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. In *Proc. of the 17th ACM Conference on Information and Knowledge Management, CIKM*, pages 449–458, New York, NY, USA, 2008. ACM.

[38] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition).* Wiley-Interscience, 2000.

[39] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine*, 17(3):37–54, 1996.

[40] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel. Analysis of geographic queries in a search engine log. In *Proc. of the 1th International Workshop on Location and the Web, LOCWEB*, pages 49–56, New York, NY, USA, 2008. ACM.

[41] Z. Ghahramani. *Advanced Lectures on Machine Learning*, chapter Unsupervised Learning. Springer Verlag, 2004.

[42] C. Grimes, D. Tang, and D. M. Russell. Query logs alone are not enough. In *Proc. of Query Log Analysis: Social and Technological Challenges Workshop at 16th International Conference on WWW*, May 2007.

[43] M. A. Hall and L. A. Smith. Practical feature subset selection for machine learning. In *Proc. of the 21th Australasian Computer Science Conference, ACSC*, pages 181 – 191, Berlin, February 1998. Springer.

[44] M. Hegland. The apriori algorithm – a tutorial. *Mathematics and Computation in Imaging Scince and Information Processing*, 9:209 – 262, March 2005.

[45] M. R. Herrera, E. S. de Moura, M. Cristo, T. P. Silva, and A. S. da Silva. Exploring features for the automatic identification of user goals in web search. *Information Processing and Management*, 46:131–142, March 2010.

[46] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of the 15th Conference on Uncertainty in Artificial Intelligence, UAI*, pages 289–29, San Francisco, CA, 1999. Morgan Kaufmann.

[47] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, International Computer Science Insitute, Berkeley, CA, 1998.

[48] G. Hripcsak and A. S. Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association, JAMIA*, 12(3):296–298, May-Jun 2005.

[49] J. Hu, G. Wang, F. Lochovsky, J. T. Sun, and Z. Chen. Understanding user's query intent with wikipedia. In *Proc. of the 18th International Conference on WWW*, pages 471–480, New York, USA, 2009. ACM.

[50] E. B. Hunt. *Artificial Intelligence*. Academic Press, New York, 1975.

[51] B. J. Jansen and D. L. Booth. Classifying web queries by topic and user intent. In *Proc. of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems*, pages 4285–4290, New York, NY, USA, 2010. ACM.

[52] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proc. of the 16th International Conference on WWW*, pages 1149–1150. ACM Press, 2007.

[53] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing and Management*, 44(3):1251–1266, 2008.

[54] B. J. Jansen and A. Spink. An analysis of web searching by European AlltheWeb.com users. *Information Processing and Management*, 41(2):361–381, 2005.

[55] B. J. Jansen and A. Spink. How are we searching the world wide web?: a comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[56] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[57] B. J. Jansen, A. Spink, and I. Taksa. *Handbook of Research on Web Log Analysis*. Idea Group Inc., 2008.

[58] X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *Proc. of the 10th ACM Conference on knowledge Discovery and Data Mining, SIGKDD*, pages 197–205, New York, NY, USA, 2004. ACM Press.

[59] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proc. of the 26th International ACM SIGIR Conference*, pages 64–71, New York, NY, USA, 2003. ACM Press.

[60] G. Karypis. Cluto a clustering toolkit. Technical Report 02-017, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, 2003.

[61] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink. Classifying the user intent of web queries using k-means clustering. *Internet Research*, 5(20):563–581, 2010.

[62] T. Kohonen. *Self organization and associative memory*, volume 8 of *Springer Series in Information Sciences*. Springer Verlag, Berlin, 1988.

[63] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 1998.

[64] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):16, March 1977.

[65] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. In *Proc. of the 14th International Conference on WWW*, pages 391–400, New York, NY, USA, 2005. ACM Press.

[66] C. Lin, G.-R. Xue, H.-J. Zeng, and Y. Yu. Using probabilistic latent semantic analysis for personalized web search. In *Proc. of Web Technologies Research and Development*, pages 707–717, Berlin Heidelberg, 2005. Springer-Verlag GmbH.

[67] Y. Liu, M. Zhang, L. Ru, and S. Ma. Automatic query type identification based on click through information. In *Proc. of the Alliance of Information and Referral Systems Conference, AIRS*, pages 593–600, 2006.

[68] D. J. C. MacKay. *Information theory, inference, and learning algorithms*, chapter An Example Inference Task: Clustering. Cambridge University Press, 2003.

[69] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, 1967. University of California Press.

[70] M. Meila. An accelerated Chow and Liu algorithm: Fitting tree distributions to high-dimensional sparse data. In *Proc. of the 16th International Conference on Machine Learning, ICML*, pages 249–57. MIT, 1999.

[71] M. Mendoza and R. Baeza-Yates. A web search analysis considering the intention behind queries. In *Proc. of the 6th Latin American Web Conference*, pages 66–74, Washington, DC, USA, 2008. IEEE Computer Society.

[72] M. Mendoza and J. Zamora. Building decision trees to identify the intent of a user query. In *Proc. of the 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems:*

Part I, KES '09, pages 285–292, Berlin, Heidelberg, 2009. Springer-Verlag.

[73] B. Mobasher. *Practical Handbook of Internet Computing*, chapter Web Usage Mining and Personalization. Chapman Hall & CRC Press, Baton Rouge, 2005.

[74] D. F. Nettleton and R. Baeza-Yates. Busqueda de información en la Web: técnicas para la agrupación y selección de las consultas y sus resultados. In *CEDI LFSC*, Granada, Spain, 2005.

[75] V. B. Nguyen and K. Min-Yen. Functional faceted web query analysis. In *Proc. of Query Log Analysis: Social and Technological Challenges Workshop at 16th International Conference on WWW*, Alberta, Canada, May 2007.

[76] A. Ntoulas, J. Cho, and C. Olston. What's new on the Web?: the evolution of the web from a search engine perspective. In *Proc. of the 13th International Conference on WWW*, pages 1–12. ACM Press, 2004.

[77] E. Piek-Vossen. *EuroWordNet. A multilingual database with lexical semantic networks*. Kluwer Academic, 1998.

[78] E. Pitler and K. Church. Using word-sense disambiguation methods to classify web queries by intent. In *Proc. of the ACM Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1428–1436, Stroudsburg, PA, USA, 2009. ACM.

[79] B. Poblete. *Query–Based Data Mining for the Web*. PhD thesis, Pompeu Fabra University, Barcelona, Spain, 2009.

[80] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[81] J. J. Randolph. Free-marginal multirater Kappa: An alternative to Fleiss fixed-marginal multirater Kappa. Paper presented at the Joensuu University Learning and Instruction Symposium. Joensuu, Finland, October 2005.

[82] D. E. Rose and D. Levinson. Understanding user goals in Web search. In *Proc. of the 14th International Conference on WWW*, pages 13–19. ACM Press, 2004.

[83] A. Schein, A. Popescul, and L. Ungar. Pennaspect: A two-way aspect model implementation. Technical Report MS-CIS-01-25, Department of Computer and Information Science, The University of Pennsylvania.

[84] A. Schein, A. Popescul, L. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proc. of the 25th International ACM SIGIR Conference*, pages 253–260, New York, NY, USA, 2002. ACM.

[85] C. Silverstein, H. Marais, M. Henzinger, and M. M. Analysis of a very large web search engine query log. In *SIGIR Forum*, pages 6–12, 1999.

[86] J. Sim and C. C. Wright. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *PHYS THER*, 85(3):257–268, 2005.

[87] V. Sindhwani and S. S. Keerthi. Large scale semi-supervised linear svms. In *Proc. of the 29th International ACM SIGIR Conference*, pages 477–484. ACM Press, 2006.

[88] N. Søndberg-Madsen, C. Thomsen, and J. M. Pena. Unsupervised feature subset selection. In *Proc. of the Workshop on Probabilistic Graphical Models for Classification*, pages 71–82, 2003.

[89] M. Speretta and S. Gauch. Personalized search based on user search histories. In *Proc. of the 2005 IEEE/WIC/ACM International Conference*

on Web Intelligence, pages 622–628, Washington, DC, USA, 2005. IEEE Computer Society.

[90] A. Spink and B. J. Jansen. A study of web search trends. *Webology*, 1(2), December 2004.

[91] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–109, 2002.

[92] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology, JASIST*, 52(3):226–234, 2001.

[93] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. Web usage mining: discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, January 2000.

[94] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proc. of Text Mining Workshop at 6th ACM SIGKDD Conference on knowledge Discovery and Data Mining*, 2000.

[95] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proc. of the 13th International Conference on WWW*, pages 675–684, New York, NY, USA, 2004. ACM Press.

[96] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*, chapter Cluster Analysis: Basic Concepts and Algorithms. Addison–Wesley, 2006.

[97] V. Tan, A. Anandkumar, L. Tong, and A. S. Willsky. A large-deviation analysis for the maximum likelihood learning of tree structures, 2009.

[98] J. Teevan, S. T. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction, TOCHI*, 17:4:1–4:31, April 2010.

[99] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proc. of the 31th International ACM SIGIR Conference*, pages 163–170, New York, NY, USA, 2008. ACM.

[100] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, NY, 1998.

[101] R. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. *Proc. of the 2nd ACM International Conference on Web Search and Data Mining, WSDM*, Jan 2009.

[102] I. H. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques*. Series in data management systems. Morgan Kaufman, Amsterdam, 2 edition, 2005.

[103] Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data mining and knowledge discovery*, 10:141–168, March 2005.