Honours Year Project Report

Restrictor Queries in Library Catalog Queries

By
Kalpana Kumar

Department of Computer Science
School of Computing
National University of Singapore
2006/2007

Project No: H079230
Advisor: Assoc Prof Kan Min-Yen
Deliverables:
Report: 2 Volumes
Acknowledgements

I would like to take this opportunity to thank my supervisor, Dr Kan Min-Yen, without who’s help and guidance, this project would not have been possible. Thanks are also due to the members of the WING group for their constant feedback throughout the project. Finally, a very special thanks to my family and close friends who supported my all through this project and gave me the confidence to see it to the finish.
Abstract

One of the most common challenges faced in designing automated systems for online library search is determining the user’s intention during a subject search. We propose the idea of a restrictor query as one way of dealing with this problem. The primary contribution of this thesis is to analyse subject queries and suggest a formal definition for restrictor queries and a comprehensive classification scheme for restrictors. A formal evaluation of the proposed definitions is conducted with human participants, and the tests show positive results in terms of replicability. Finally, we propose an automatic way of identifying restrictor queries and classifying restrictors by adopting the Naïve Bayes approach. This study is meant to serve as a first step to analysing subject queries according to their semantic meaning and returning the most relevant results based on this analysis, thus enriching the user search experience.
Table of Contents

Acknowledgements i

Abstract ii

Table of Contents iii

List of Figures v

List of Tables vii

Chapter 1: Introduction
  1.1 Motivation 1
  1.2 Objectives of this Project 4
  1.3 Project Contributions 5
  1.4 Structure of this Report 5

Chapter 2: Related Work
  2.1 Taxonomies for Web and OPAC query classification 6
  2.2 Systems for Automatic Web and OPAC query classification 8

Chapter 3: Restrictor query:
  Formal Definition and Classification
    3.1 Methodology used to derive the definition of a restrictor query
       3.1.1 Examination of existing classifications for Adjuncts and Adjectives 10
       3.1.2 Analysis of hyponym structure in WordNet 12
       3.1.3 Analysis of queries from LINC 13
    3.2 Formal Definition of a restrictor query 14
    3.3 Classification of restrictors in restrictor queries 15
    3.4 Exceptions of Restrictor Queries
       3.4.1 No obvious core term 18
       3.4.2 A restrictor modifying another restrictor 19
       3.4.3 More than one core term 19

Chapter 4: Evaluation of Definitions-
Experimental Methodology

4.1 Pre-Experiment Preparation
  4.1.1 Selection of Test Participants 20
  4.1.2 Preparation of Test Data 21
  4.1.3 Assessment Modes 21

4.2 Experimental Methodology
  4.2.1 Understanding Key Terminologies 21
  4.2.2 Reflecting knowledge in Experimental Tasks 22

Chapter 5: Evaluation of Definitions-
Results and Analysis

5.1 Judgement of a restrictor versus a non-restric tor query
  5.1.1 Variation between average responses of participants 24
  5.1.2 Agreement between participants and assessor 26

5.2 Selection of core term 28

5.3 Selection of restrictors
  5.3.1 Agreement between participants and assessor 31
  5.3.2 Relation between class of restrictors and linguistic behaviour 32

5.4 Level of Confidence of choice 32

Chapter 6: Automatic Classification

6.1 Overview and objectives of the proposed algorithm 35
6.2 Naïve Bayes Classifier 36
6.3 Features used in our algorithm
  6.3.1 Linguistic Features 37
  6.3.2 Distributional Features 38
  6.3.3 Semantic Features 39

6.4 Execution of Algorithm
  6.4.1 Query Classification 41
  6.4.2 Restrictor Identification 42
  6.4.3 Restrictor Classification 43

6.5 Scaling the Algorithm 44
6.6 A real-life example 45

Chapter 7: Conclusion

7.1 Limitations of our study 48
7.2 Areas for future research 49
7.3 Conclusion 49

References 50

Appendices 52
## List of Figures

| Figure 1.1 | Screen-shot of search results for the query ‘Economic History of Singapore’ as presented by the NUS LINC | 3 |
| Figure 1.2 | Screen-shot for search results for an ideal system that can interpret the meaning of a query and provide categorized results | 4 |
| Figure 3.1 | Formal definition of a restrictor query | 14 |
| Figure 3.2 | Overview of the proposed classification of restrictors | 16 |
| Figure 4.1 | Informal Definition of a restrictor query | 22 |
| Figure 4.2 | An example where the participants had to judge if the query was a restrictor query | 22 |
| Figure 4.3 | An example where the participant had to classify the base term and restrictors of a restrictor query | 23 |
| Figure 4.4 | Example where participants had to choose their level of confidence | 23 |
| Figure 5.1 | Comparison between participants’ and assessor’s agreement over restrictor queries | 26 |
| Figure 5.2 | Distribution of false-negatives according to the various restrictor classes | 30 |
| Figure 5.3 | Distribution of Confidence levels of participants | 33 |
| Figure 5.4 | Average confidence levels for various types of queries | 34 |
| Figure 6.1 | Overview of the work-flow of the proposed algorithm | 35 |
| Figure 6.2 | Call Numbers as seen in LINC | 39 |
| Figure 6.3 | Flow-chart to show Query Classification at Runtime | 41 |
| Figure 6.4 | Flow Chart to show Restrictor identification at runtime | 42 |
| Figure 6.5 | Flow-chart showing the procedure of Restrictor classification at runtime | 43 |
| Figure 6.6  | Query identification | 45 |
| Figure 6.7  | Three iterations to compare the terms of the query, after which literature is determined as the core term | 46 |
| Figure 6.8  | Classification of the restrictors into the desired classes | 47 |
List of Tables

Table 3.1  Artificially created list of queries from LINC  
Table 3.2  Separation of restrictor and non-restrictor queries for the queries presented in Table 3.1  
Table 3.3  Restrictors for the queries classified as restrictor queries  
Table 5.1  Summary of Results for Task 1  
Table 5.2  ANOVA analysis of results for Task 1  
Table 5.3  F and p values for the five individual sets of data  
Table 5.4  Some known-item queries that were marked as restrictor queries  
Table 5.5  Comparison of responses for queries with different parts-of-speech  
Table 5.6  Categorization of core term agreement  
Table 5.7  Results for core term agreement between assessor and participant  
Table 5.8  Some queries in which there was a disagreement between participants and assessor  
Table 5.9  Contingency table to show the distribution of terms  
Table 5.10  Summary of Levels of Confidence of the participants  
Table 6.1  Description of exhaustive search method to scale the proposed algorithm
Chapter 1

Introduction

1.1 Motivation
Revolutionary advancements in technology over the past 50 years have transformed the library-user relationship, not only in how library records are presented to users, but also in how users interact with the library. With the advent of the internet, the traditional notion of a library itself has undergone significant change, from a physical entity where information can be obtained in one place, to modern online versions which offer the user greater convenience and mobility. Work to further improve such systems now focus on implementing algorithms and designing features that can automate the entire search process to obtain the most relevant search results.

Yet one of the most common challenges faced in designing such automated systems is aligning the user’s objectives with the capabilities of the system. As pointed out by Slone (2000), a user hopes to find an answer or the item that meets his/her information need, but current systems, by far and large, have no way of inferring the user’s intention and rely on simple keyword matching. This gap between understanding one’s own intention and the actual expression as a search query has been termed the Anomalous State of Knowledge (ASK) (Belkin, 1980) and continues to be a problem today. Unskilled users of such search systems have been found to employ generic search terms to express their objectives, leading to information overload (Nordie, 1999). Differences in the interpretation of queries between humans and automated systems can often lead to poor search results.
Much research has been done to understand user goals based on how users modify their search terms during the course of their interaction with the automated system (Hert, 1996; Nordie, 1999; Slone, 2000; Case, 2006). Yet, few studies have made attempts to actually model this behaviour in automated systems. Kan and Poo (2005) have suggested an automated way of distinguishing known-item queries (searches for a particular title or author) from unknown-item or subject queries (searches for resources on a particular subject) and area queries (searches for particular sections of the library). However, this system deals only with determining the user’s intention when he has a specific title or author in mind. Establishing what the user hopes to find when he enters a subject query is still a major hurdle faced by many Online Public Access Catalogues (OPACs) today.

The primary contribution of this thesis is to analyse subject queries and suggest one way of giving relevant results when a user enters such a query into the system. Buckley (2004), in his analysis of why information systems fail, has concluded that being able to understand the semantics of a subject well enough to determine its important aspects is crucial for searches. This motivates us to propose the notion of a restrictor query, as a subject query which is composed of one generic term which captures the core meaning of the query, and other terms which limit the scope of this core term, or make it more specific (A more formal definition of a restrictor query is given in Chapter 3).

At the National University of Singapore (NUS), information on library resources is maintained by an online catalogue known as the Library INtegrated Catalogue (LINC). An initial survey of subject queries from simple keyword searches in LINC revealed that the semantic meaning of a query can be identified by determining a core term which gives the query its fundamental meaning, and other terms which further narrow down this core term. Thus, one approach to finding the most relevant results for users would be to analyse the semantic meaning of the search query by categorizing the query based on the core term and its relationship to its modifiers, and ranking the titles according to this relationship. This will enable users to have a focused selection of titles to choose from, thus obtaining relevant results in a shorter span of time and making their search experience more fruitful.
As an example, for a subject query such as ‘Economic History of Singapore’ entered via the keyword function of OPACs, most search engines would present the results to the user as one comprehensive list as shown in Figure 1.1:

![Figure 1.1 Screen-shot of search results for the query ‘Economic History of Singapore’ as presented by the NUS LINC](image)

However, suppose the system had the ability to analyse the semantic meaning of the query and deduce its core and modifying terms, it then might be able to make suggestions for only those sections of the library from which the user can obtain the most appropriate resources. The user will thus see a different picture from the one portrayed in Figure 1.1. Figure 1.2 below is one possible representation of how search results might be ranked given the system is able to conduct a semantic analysis of the query.
Figure 1.2  Screen-shot for search results for an ideal system that can interpret the meaning of a query and provide categorized results

Here, the query has been analysed according to its core meaning, and not only is the user given information about which sections in the library can give him the most relevant resources, but he is also able to view a categorised list of only those titles that relate to the various aspects of the query. In the above example, for the query ‘Economic History of Singapore’, no longer does he get just resources that match the terms ‘Economic’, ‘History’ and ‘Singapore’ in the title name, but he now gets appropriate titles that have been classified according to the generic topic of ‘Economic History’, and more specifically, those that refer to the ‘Economic History of Singapore’.

1.2  Objectives of this Project

The objectives of this project are:

- To conduct an in-depth analysis of search queries from LINC and formalize a definition for a restrictor query;
- To determine a classification scheme for various restrictors. Knowing what type of restrictor exists in the query will enable the system to decide which subject areas might be most relevant to the user;
• To determine whether the proposed definition and classification scheme is reproducible through a series of experiments with human participants;
• To propose an algorithm to automatically detect restrictor queries and classify the restrictors according to the proposed classification scheme.

1.3 Project Contributions
The definitions and algorithm proposed in this project are not limited to the domain of an academic OPAC. These can be applied to analyse queries in public libraries, as well as in other forms of search, such as web search engines. This formal definition and classification algorithm serve as the first step towards integrating query analysis with the way in which data is organised in libraries, thus enhancing the user experience, and play a significant role in incorporating human behaviour further with the field of text recognition.

1.4 Structure of this Report
The rest of the discussion is divided into six distinct chapters which cover the theoretical research done, the definitions proposed, the empirical tests carried out to verify the reproducibility of the definitions and the proposed algorithm that deals with automating the classification process. Chapter 2 reviews the literature on taxonomies for classification, both in the Web and OPAC domain, and looks at techniques previously proposed for automating the detection of such classification schemes. The formal definition of restrictor queries and the classification scheme are discussed in Chapter 3. Chapter 4 gives a broad overview of the various tests conducted to judge human agreement with the proposed definitions. A detailed discussion of the results obtained is conducted in Chapter 5 in order to assess if the definitions are replicable and to gather evidence for justifying features proposed in our algorithm. The algorithm to automatically classify restrictor queries is then proposed in Chapter 6, along with real-life examples that illustrate the working of the algorithm. Finally, Chapter 7 acknowledges the limitations of the study and suggests directions for further research, before drawing conclusions based on the preceding chapters and discussing technical and academic implications of this study.
Chapter 2

Related Work

This chapter reviews related work in the field of query classification, focusing on taxonomies introduced for both Web and OPAC query classifications. Parallels can be drawn between web and OPAC classification schemes, hence it is important to review both these domains together. The next section introduces systems that have been built to deal with automatic classification schemes in both domains, reflecting on their strengths and limitations and ultimately reinforcing the motivation for this project.

2.1 Taxonomies for Web and OPAC query classification

Identifying the user’s intention behind a particular search has been recognized as one of the most important steps in automating the information retrieval process. In web searches, identifying the user’s intentions behind each search is an extensive and complicated task. Broder (2002) suggested the first taxonomy for web queries, based on the following three classes of user goals:

- **Navigational**, where the user intended to obtain a specific website that he had in mind.
- **Informational**, where the user was looking for any information on a particular topic.
- **Transactional**, where the user intended to make use of a particular service offered by a webpage (for example, download software from that page).

Rose and Levinson (2004) delved further into the subject of taxonomies for web queries, and proposed a more in-depth taxonomy comprising **Navigational**, **Informational** and
Resource search goals, with sub-categories for each class. They also presented a comparison of various web query classification schemes and concluded that identifying user goals without any additional information, such as user-click history, was almost as effective as in the case when external information was used.

As opposed to web searches, OPAC searches take place in a closed environment. That is, users search the OPAC only for material related to certain subjects that may have to do with their curriculum projects, for example, and are restricted to only obtaining resources present in the library’s database as search results. Yet some parallels can be drawn between some classes of queries that have been proposed for both domains.

An attempt to understand user goals in OPAC searches has been made by Hert (1996), in which the author proposes four different intentions that the user might have when conducting a search on an OPAC. These are:

- Searching for specific (known) entities (a resource of the same name that already exists)
- Searching for unknown entities (a resource on a particular subject)
- Searching for information on specific entities, such as a journal or a video clip etc.
- Searching for information, without specification of numbers or types of entities.

The author also identified a set of situational elements specific to each user goal, and discussed the implications of her findings on OPAC design.

A similar distinction was made by Slone (2000), who proposed three classes of searches that users of OPACs perform:

- Known-item searches, in which the user is looking for a particular title or author.
- Unknown-item searches, in which the user is looking to obtain resources to address a particular issue, or resolve a problem, and
- Area searches, in which the user is trying to find a particular area or section of the library.
The author further analysed the different strategies adopted by OPAC users in obtaining the relevant information for each of these types of searches in a bid to provide an area for future research for librarians and interface designers.

A similarity can be drawn between user goals in Navigational and known-entity classes, and between Informational and unknown-entity classes. In both Navigational and known-entity searches, the user is looking for a particular resource that he has in mind, while in Informational and unknown-entity searches, the user is simply looking for information on a particular subject, or for resources that can provide this information. A third, not so obvious comparison can be made between the Resource goals of web users and the specific entity search made by OPAC users. In both instances, the user is interested in specific entities that serve a particular purpose.

Given that similarities between web and OPAC query classifications exist, it may be possible to deduce that strategies employed in automated systems for query classification in one domain may be adapted for classification schemes proposed in the other. In the next section, we take a look at some of these automated classification techniques and highlight their strengths and weaknesses.

2.2 Systems for Automatic Web and OPAC query classification

Based on the taxonomy suggested by Broder, Kang and Kim (2003) proposed an automatic classification scheme for web queries, which incorporated both features of the query itself (such as parts-of-speech(POS)) as well as other information about the query (such as the URL information). They first classified queries according to three different tasks: those that performed the topic relevance task, the homepage finding task and the service finding task. They then adopted a different strategy for each task, based on factors such as the difference of distribution, mutual information, the usage rate as anchor texts, and the POS information.

Lee, Liu, Cho (2005) improved on the scheme proposed by Kang and Kim to develop a system for automatically determining user goals using features such as click behaviour
and anchor-link distribution that was accurately able to determine user goals in 90% of Navigational and Informational queries. Although both these approaches yield encouraging results with respect to identifying user goals, they rely extensively on external information such as the click history of the user itself to deduce his search intention, without much of an analysis of the semantics of the query itself.

The area of automatic classification for OPAC queries is still relatively unexplored. One approach proposed by Kan and Poo (2005) works around the problem of too much dependence on external information as seen in the systems for web query classification by proposing methods that treat each query as a separate entity without reliance on external factors. The authors suggested an automated way of distinguishing known-item queries from unknown-item and area queries. The first part of their research involved identifying traits specific to known-item queries that helped distinguish them from other types of queries. They then employed methods in machine learning, language modeling and machine translation evaluation metrics to decide whether a user query qualified as a known-item query, and whether the query returned the particular resource sought. The system proposed by the authors had an 80% and 95% correlation to human performance for both tasks respectively. The methodology of identifying traits specific to known-item queries forms the motivation for part of our methodology to formalize a definition for restrictor queries. This paper also provides ideas for systems to intelligently handle queries without using external data and forms the basis for the ideas proposed in our system. Yet, this system deals only with determining the user’s intention when he has a specific title or author in mind. In most OPACs, searches for unknown entities still account for almost 70 percent of all searches. Thus, establishing the user’s intention during a subject search can be perceived to have an impact on a significant number of searches conducted. As this problem is yet to be addressed in OPACs today, it forms the basis for the motivation for this project.

---

1 Assumption made according to analysis of query logs obtained from LINC
Chapter 3

Restrictor query: Formal Definition and Classification

In this chapter, we propose a formal definition of a restrictor query, after examining the various approaches that were employed in constructing this definition.

3.1 Methodology used to derive the definition of a restrictor query
A variety of methods were adopted in determining a formal definition for a restrictor query and a classification scheme for restrictors. These methods are briefly discussed below:

3.1.1 Examination of existing classifications for Adjuncts and Adjectives
Ideas for a good classification scheme for restrictors stemmed from looking at existing schemes for classifying certain linguistic patterns specific to the English language that are of relevance to this project. Of particular interest are the argument-adjunct distinction, the classification of adjuncts, and the classification of adjectives.

A. The argument-adjunct distinction and classification of adjuncts
In linguistics, arguments and adjuncts form different parts of a sentence and vary in terms of their functional value, as well as in the way in which they are interpreted (Grimshaw, 1990). According to the Dictionary of Linguistics and Phonetics, an argument is defined as a primary entity in a sentence construction that bears a direct semantic relation to the head term of the sentence, while an adjunct is defined as an optional, secondary element
in a sentence construction which, when removed, does not affect the structure identity of the rest of the sentence.

While both arguments and adjuncts are important for sentence structure, and help establish the circumstances under which the state or action denoted by the head term (the main phrase of the sentence, which could either be a noun phrase or a verb) takes place, adjuncts tend to be more independent of the head term. Merlo and Ferrer (2005) and Grimshaw (1990) have noted some key distinctions between arguments and adjuncts, by which they can be easily discerned from each other. These rules are as follows:

• Functionally, an argument describes a role determined by its associated head, whereas an adjunct propagates a property of the head that is independent of the associated head’s state or functionality.

• A phrase can be classified as an argument if its meaning depends mainly on its associated head and as an adjunct if its meaning does not change in the context of different heads.

• An adjunct can co-occur with a relatively broad range of heads, while an argument, due to its more dependent nature, can only occur with a semantically restricted domain of head words.

• Adjuncts are independent entities, that is, removing an adjunct from a sentence will not alter the structure of the sentence.

Merlo and Ferrer also proposed an automatic way of distinguishing arguments from adjuncts by using diagnostics such as head dependence, optionality, iterativity, ordering, corpular paraphrase and deverbal nominalization.

While we do not know of a classification for arguments, the classification of adjuncts has formed the basis for various in-depth studies. Adjuncts can be classified as:

• *Temporal* adjuncts, which establish when, for how long or how often a state or action happened or existed.

• *Locative* adjuncts, which establish where, to where or from where a state or action happened or existed.
• **Modicative** adjuncts, which establish how the action happened or the state that existed when the action happened.
• **Causal** adjuncts, which establish the reason for, or purpose of, an action or state.
• **Instrumental** adjuncts, which establish the instrument of the action.
• **Agentive** adjuncts, which establish the agent of the action.
• **Conditional** adjuncts, which establish the condition in which a sentence becomes true.
• **Concessive** adjuncts, which establish the contrary circumstances.

B. **Classification of adjectives**
While classes of adjectives can be commonly found in most linguistics text books, of particular interest is the distinction proposed by Willners (1997), in which the author partitions all adjectives into three main classes:

• Descriptive adjectives that assign the value of an attribute to a noun.
• Relational adjectives that denote an association with the noun.
• Reference-modifying adjectives, which, as the name suggests, modify the reference to a noun. This class of adjectives can only occur in attributive positions and the nouns they modify generally denote a function or a social relation.

3.1.2 **Analysis of hypernym structure in WordNet**
WordNet (http://wordnet.princeton.edu/) is a lexical reference system that contains nouns, verbs, adjectives and adverbs that are grouped together into sets of cognitive synonyms, each of which expresses a distinct concept. These sets are then linked by means of conceptual-semantic and lexical relations. WordNet’s structure makes it a useful tool in Computational Linguistics and Natural Language Processing.

The most significant feature of WordNet is its semantic organization. Of particular interest to this project is the ‘inherited hypernym’ representation, which provides the generic term used to denote a whole class of specific instances. This feature was of significant use in determining broad class names for restrictors that were similar in nature, as well as in deciding which types of restrictors could be combined under the
same class name, as the generic class gave an idea of which classes of words were essentially similar in their broad sense.

3.1.3 Analysis of queries from LINC

Close to 700,000 library queries generated over two years at the NUS LINC were obtained. Of these, about 12,000 queries were manually analysed to determine which queries might form part of the domain of restrictor queries. Table 3.1 shows an artificially created list of queries as seen in a LINC query log. This list will serve as a running example to illustrate the definitions proposed.

| American Journal of Psychology                        |
| Material properties of silicon                        |
| Festivals in Thailand                                 |
| Children                                               |
| Economic history of Singapore, 1812-1963               |
| A pound of flesh                                      |
| Face recognition                                      |
| (calculus) and (engineering)                          |
| European Literature in the eighteenth century         |
| Rituals in Hinduism                                   |
| system modeling                                       |
| Slow chemical reactions                               |
| ‘language attitudes’ and ‘students’                    |
| Marine biology of ocean species                       |
| Accounting fraud and political law suites in Ghana    |

Table 3.1 Artificially created list of queries from LINC

Manual observation of the queries gave rise to some heuristic rules concerning restrictor queries as listed below:

- A restrictor query should be made up of at least two words.
- Restrictor queries cannot be known-item queries.
• Restrictor queries should have at least two parts, of which one is a generic term, and the other is a term that modifies the scope of the generic term, or makes it more specific. These two parts can be explicitly separated by a conjunction or a preposition, or implicitly separated, as in the case of an adjective attached to a noun. A restrictor query must have at least one generic term, and at least one modifying term.

• Generic terms of restrictor queries tend to be either noun or verb phrases, while the modifying terms usually tend to be adjectives, adverbs, numerals or proper nouns. Semantically, many modifying terms represent geographical locations, dates or quantities.

The rules presented above form the basis for the linguistic features adapted in the proposed algorithm, and will be discussed in further detail in Chapter 6.

3.2 Formal Definition of a restrictor query

Following the initial analysis of query logs, hyponym structures and the examination of existing classifications, a formal definition for a restrictor query was proposed as follows:

In the domain of an OPAC, a query $q$ is a restrictor query, if $q$ is a subject query, and consists of at least one core term (comprising the argument and its associated head word or phrase) and at least one adjunct which limits the scope of the core term.

Figure 3.1 Formal definition of a restrictor query

Based on this formal definition and the heuristic rules discussed in section 3.1.3, we can determine which queries from table 3.1 form part of the domain of restrictor queries. Table 3.2 shows this division, with reasons for why the non-restrictor queries are classified as such.
<table>
<thead>
<tr>
<th>Restrictor Queries</th>
<th>Non-Restrictor Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Material properties of silicon</td>
<td>American Journal of Psychology</td>
</tr>
<tr>
<td></td>
<td>• Not a restrictor query as it is the title of a journal (i.e. a known-item query).</td>
</tr>
<tr>
<td>Festivals in Thailand</td>
<td>Children</td>
</tr>
<tr>
<td></td>
<td>• Not a restrictor query as it consists of only one word</td>
</tr>
<tr>
<td>Economic history of Singapore, 1812-1963</td>
<td>Face recognition</td>
</tr>
<tr>
<td></td>
<td>• Not a restrictor query, as the terms ‘face’ and ‘recognition’ cannot be split into two distinct terms. In this case, ‘face’ functions as an argument, and thus cannot be separated from ‘recognition’ to form a restrictor.</td>
</tr>
<tr>
<td>A pound of flesh (calculus) and (engineering)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Not a restrictor query, as neither calculus, nor engineering limits the scope of the other term.</td>
</tr>
<tr>
<td>European Literature in the eighteenth century</td>
<td>system modeling</td>
</tr>
<tr>
<td></td>
<td>• Not a restrictor query, as it does not contain a clearly defined core term and restrictor.</td>
</tr>
<tr>
<td>Rituals in Hinduism</td>
<td></td>
</tr>
<tr>
<td>Slow chemical reactions</td>
<td></td>
</tr>
<tr>
<td>‘language attitudes’ and ‘students’</td>
<td></td>
</tr>
<tr>
<td>Marine biology of ocean species</td>
<td></td>
</tr>
<tr>
<td>Accounting fraud and political law suites in Ghana</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2 Separation of restrictor and non-restrictor queries for the queries presented in Table 3.1

3.3 Classification of restrictors in restrictor queries

A single restrictor query can have multiple restrictors that limit the scope of the core term by various dimensions. Hence, it is necessary to have a classification of the various types of possible restrictors. A good classification of restrictors will ultimately help in deciding which sections of the library might provide the most relevant results to the user. Figure 3.2 gives an overview of the proposed classification scheme.
The proposed classification scheme comprises four main classes of restrictors, all of which restrict queries in the domain of subject searches. An explanation of each of these classes is as follows:

A. **Locative restrictors**
Locative restrictors modify the core term by the place of occurrence.

B. **Temporal restrictors**
Temporal restrictors limit the core term by the time or duration of occurrence.

C. **Quantitative restrictors**
Quantitative restrictors limit the scope of the core term by a measurable amount. (*How much* of the core term is in question?).
D. Qualitative restrictors

Qualitative restrictors modify the core term by a particular quality of that term. This class can be further sub-divided into:

- **Descriptive restrictors** that denote *the type* of the core term. This class of restrictors usually comprises traits that are generic across most core terms.

- **Aspectual restrictors** that limit the scope of the core term by a particular characteristic of that term which is not descriptive. Here, only one particular body of information about the core term is desired, hence this class of restrictors denotes traits that are specific for only certain types of core terms.

For queries that have been classified as restrictor queries in Table 3.2, we can now classify their restrictors as follows:

<table>
<thead>
<tr>
<th>Restrictor Queries</th>
<th>Restrictors</th>
<th>Reasons for classification</th>
</tr>
</thead>
</table>
| Material properties of silicon | Core term: silicon  
Aspectual restrictor: Material properties | *Material properties* limits the scope of the core term, *silicon* to just one aspect without actually describing silicon |
| Festivals in Thailand | Core term: Festivals  
Locative restrictor: Thailand | *Thailand* denotes the place of occurrence of festivals. |
| Economic history of Singapore, 1812-1963 | Core term: history  
Descriptive restrictor: economic  
Locative restrictor: Singapore  
Temporal restrictor: 1812-1963 | • *Economic* denotes the type of *history* that is in question  
• *Singapore* denotes the place of occurrence of the core term, *history*  
• *1812-1963* denotes the time-frame of occurrence of the core term, *history*. |
| A pound of flesh | Core term: flesh  
Quantitative restrictor: pound | *Pound* denotes the amount of *flesh* that is in question |
| European Literature in the eighteenth century | Core term: literature  
Locative restrictor: European | • *European* denotes the place of
Temporal restrictor: eighteenth century

Eighteenth century limits the core term, literature by the time during which it occurred.

### Table 3.3 Restrictors for the queries classified as restrictor queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Restrictor Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rituals in Hinduism</td>
<td>Temporal restrictor: *</td>
</tr>
<tr>
<td>Slow chemical reactions</td>
<td>Core term: chemical reactions  Descriptive restrictor: slow  Slow describes the speed of the core term, chemical reactions</td>
</tr>
<tr>
<td>‘language attitudes’ and ‘students’</td>
<td>Core term: students  Aspectual restrictor: language attitudes  Language attitudes refers to just one aspect of the core term, students without actually describing the students</td>
</tr>
<tr>
<td>Marine biology of ocean species</td>
<td>*</td>
</tr>
<tr>
<td>Accounting fraud and political law suites in Ghana</td>
<td>*</td>
</tr>
</tbody>
</table>

Notice that some examples have been left intentionally blank and marked with a star. These examples will serve to highlight certain exceptions in the next section.

### 3.4 Exceptions of Restrictor Queries

Most queries have just one obvious core term and other modifying terms that limit only this core term. There are, however, cases when this rule is not observed. We highlight some of these cases here.

#### 3.4.1 No obvious core term

In some cases, it is not always clear which term forms the core term and which term forms the restrictor. When one term is chosen as a core term, the other term might fall under one class of restrictors. Yet, when the other term is chosen as the core term, the first term might fall under yet another class of restrictors. An example of such a case is in the query ‘Rituals in Hinduism’. When rituals is chosen as the core term, hinduism forms a descriptive restrictor, however when hinduism is chosen as the core term, rituals forms
the aspectual restrictor. Using only parts-of-speech information, we are unable to tell which of these is a better option for a core term, hence we propose to make use of other information such as distributional and semantic knowledge in our decision. These ideas will be discussed in greater detail in chapters 5 and 6.

3.4.2 A restrictor modifying another restrictor
In some cases, one restrictor can act as a core term for another restrictor to modify. This is seen in the example ‘Marine biology of ocean species’. The term biology forms an aspectual restrictor for the core term ocean species, as it limits the scope to just one aspect of ocean species. However, the term marine is a descriptive restrictor for biology, as it denotes what type of biology is in question. Thus, the term biology forms a core term in one case, and a restrictor in the other.

3.4.3 More than one core term
Some queries, such as ‘Accounting fraud and political law suites in Ghana’ exhibit more than one core term. The terms fraud and law suites form core terms for accounting and political, respectively, which are both descriptive restrictors. Finally, Ghana acts as a locative restrictor for both core terms. Thus, not always do we find only one core term in a query.
Chapter 4

Evaluation of Definitions-
Experimental Methodology

Theoretical analysis of a given measure for properties thought desirable are assumed to act at best as a coarse filter in the assessment of the measure. A better approach would be comparisons of a set of measures based on human judgement. Budanitsky and Herst (2000) have mentioned that such evaluation amongst human participants clearly give the best assessment of the ‘goodness’ of a measure. Accordingly, the evaluation of our proposed definitions also underwent an assessment by human participants to test whether our definitions were replicable (that is, whether any person reading the definition for the first time would be able to grasp its meaning and apply it to real-life examples immediately). The methodology adopted for this experiment, including the selection of participants and the assessment modes, is described in this chapter.

4.1 Pre-Experiment Preparation

4.1.3 Selection of Test Participants
A total of 25 test participants were selected for this experiment. The participants comprised undergraduate and graduate students from the department of Computer Science at the National University of Singapore. They were selected randomly, without any prior knowledge of their qualification or academic background. This procedure ensures the results are not biased to any particular user profile.
4.1.2 Preparation of Test Data

A total of 125 queries were selected for this experiment, comprising equally of restrictor and non-restrictor queries. The queries were individually selected from the set of 12,000 queries previously analysed. Each class of restrictors is represented in at least ten queries, with some classes being present in more queries due to the presence of multiple restrictors in one query.

The entire set of 125 queries was randomly shuffled to avoid any bias that may have occurred due to the initial segregation of the queries at the analysis stage. They were then divided into five sets of 25 queries each. Each of these five sets was then distributed amongst five test participants, thus, one batch of 25 queries was judged by five different test participants independent of each other.

4.1.3 Assessment Modes

Questionnaires and discussion sessions were used as the main modes of assessment throughout this experiment, which lasted for an hour. The participants were randomly divided into two groups, and the experiment was conducted over two sessions. During the experiment, participants were allowed to ask questions to the examiner. However, discussion between participants was not permitted.

4.2 Experimental Methodology

4.2.1 Understanding Key Terminologies

The participants were first given a brief explanation of the project and the objectives of the experiment, followed by an informal definition of a restrictor query and some examples of what was and was not a restrictor query. The definition was made informal so as to make it easier for the participants to grasp the main aspects of the definition.
The informal definition used for this experiment is as shown in Figure 4.1:

A query is a restrictor query if it constitutes a base part (usually a noun or a verb) and at least one restrictor (usually an adjective, adverb or proper noun) such that:

- The query is a subject query (ie, does not contain Titles or Names of Authors)
- The restrictor limits the scope of the base term.
- Removing the restrictor from the query does not change the core meaning of the query.

![Figure 4.1 Informal Definition of a restrictor query](image)

The participants were also provided with a classification scheme for restrictors, which resembled that given in Chapter 3. They were given about ten minutes to read through these definitions, after which a discussion session followed during which they were allowed to bring up any clarifications related to the experiment that they had.

4.2.2 Reflecting knowledge in Experimental Tasks

The second part of the experiment involved a hands-on process, in which the participants were asked to assess some real life queries based on their understanding of the meaning of a restrictor query. Each participant was given a data set of 25 queries that was complied as described in the previous section and asked to perform three tasks. The participants were given 30 minutes to perform this part of the experiment.

A. Judge if the query is a restrictor query

The participants first had to decide whether the query was a restrictor query. They had to give their answers in the form of a simple ‘Yes’ or ‘No’ and had to circle their choice, as shown in the example in Figure 4.2

![Figure 4.2 An example where the participants had to judge if the query was a restrictor query](image)
B. Classifying the restrictors

For each query that the participants classified as a restrictor query, they were asked to specify the base term and the restrictors, and categorize the restrictors according to the classification scheme outlined. If the participant had indicated a ‘No’ in the previous task, he was asked to skip this task and proceed to the next task. The participants had to note down the base term and each restrictor, and classify them by using ‘L’ for locative, ‘T’ for temporal, ‘Q’ for quantitative, ‘D’ for descriptive and ‘A’ for aspectual restrictors. An example of a participant’s response for the same query as above is shown in Figure 4.3.

If Yes, the base term is: History

Indicate the restrictors and the type below:

<table>
<thead>
<tr>
<th>Restrictor</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic of Singapore</td>
<td>D</td>
</tr>
<tr>
<td>1912</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>T</td>
</tr>
</tbody>
</table>

Figure 4.3 An example where the participant had to classify the base term and restrictors of a restrictor query

C. Deciding the Level of Confidence of their choice

The final task for each query was for the participants to decide how sure they were of their choice. For this task, the participants were given five options, ranging from ‘Not sure at all’ to ‘Very sure’ and had to choose their level of confidence, as illustrated in Figure 4.4.

How sure are you of your choice?
[ ] Not sure at all  [ ] Not very sure  [ ] Neutral  [ ] Sure  [ ] Very sure

Figure 4.4 Example where participants had to choose their level of confidence

For a sample of the dataset provided to the participants, please refer to Appendix B. A detailed discussion of the results obtained is made in the next chapter.
Chapter 5

Evaluation of Definitions-
Results and Analysis

In this chapter, we present a detailed analysis of the results obtained from the experiments conducted. The analysis is meant to serve as evidence for the methodology and approaches adopted in our proposed algorithm for automatic query classification. It is also meant to serve as a basis for further academic research in the area. In this chapter, the responses of the participants are referred to as those of the ‘participant’ or ‘respondent’, while our responses are referred to as those of the ‘assessor’.

5.1 Judgement of a restrictor versus a non-restrictor query

5.1.1 Variation between average responses of participants

Table 5.1 gives a summary of the results obtained for the first task in the experiment, where the participants were asked to judge if a query was a restrictor query. The responses have been combined so as to reflect them as the work of five individuals only. The responses of five participants, each of which graded different data sets, were concatenated, so that each participant is assumed to have assessed 125 queries.

<table>
<thead>
<tr>
<th>Groups</th>
<th># queries assessed</th>
<th># queries judged as RQs</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>125</td>
<td>78</td>
<td>0.624</td>
<td>0.236516129</td>
</tr>
<tr>
<td>Participant 2</td>
<td>125</td>
<td>84</td>
<td>0.672</td>
<td>0.222193548</td>
</tr>
<tr>
<td>Participant 3</td>
<td>125</td>
<td>74</td>
<td>0.592</td>
<td>0.243483871</td>
</tr>
<tr>
<td>Participant 4</td>
<td>125</td>
<td>77</td>
<td>0.616</td>
<td>0.238451613</td>
</tr>
<tr>
<td>Participant 5</td>
<td>125</td>
<td>68</td>
<td>0.544</td>
<td>0.250064516</td>
</tr>
</tbody>
</table>

Table 5.1 Summary of Results for Task 1
These results were then analysed using a single-factor Analysis of Variance, or ANOVA (Rose, Sixth edition 2002) test, which evaluates whether there are differences between the average responses across the five users. Table 5.2 shows the results of this analysis.

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>1.0944</td>
<td>4</td>
<td>0.2736</td>
<td>1.148895</td>
<td>0.332471</td>
<td>2.386303</td>
</tr>
<tr>
<td>Within Groups</td>
<td>147.648</td>
<td>620</td>
<td>0.238142</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2 ANOVA analysis of results for Task 1

The ANOVA test uses estimates of the variance to compare average responses. The terms ‘within groups’ and ‘between groups’ refer to the variation between individual responses of one participant, and between the overall responses of different participants. Mathematically, this variation can be calculated using the $F$-value, which tests the hypothesis that the average responses of the participants are the same. This $F$-value compares the ratio of the variation within groups to the variation between groups. In order to conclude that there is no significant difference in the average responses of the participants, the $F$-value calculated must be close to 1.

The $P$-value indicates the probability that our calculated $F$-value would be obtained by chance (random error) alone. A higher $P$-value (>0.05) indicates that there is a greater likelihood that the difference in the average responses of participants is due to random factors, as is the case in our analysed results ($P = 0.33$).

A similar trend is seen in the individual results of the five sets of data assessed. The $F$-values indicate the average variation in the responses of the five individuals grading the same data set. Table 5.3 summarizes these results:

<table>
<thead>
<tr>
<th></th>
<th>F-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>1.16</td>
<td>0.33</td>
</tr>
<tr>
<td>Set 2</td>
<td>0.99</td>
<td>0.42</td>
</tr>
<tr>
<td>Set 3</td>
<td>0.94</td>
<td>0.44</td>
</tr>
<tr>
<td>Set 4</td>
<td>0.95</td>
<td>0.44</td>
</tr>
<tr>
<td>Set 5</td>
<td>1.21</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 5.3 $F$ and $p$ values for the five individual sets of data
5.1.2 Agreement between participants and assessor

Figure 5.1 given a comparison between participant and assessor responses for the agreement of restrictor queries.

![Comparison between participants' and assessor's agreement over restrictor queries](image)

While the general consensus between participants was good, as was seen in the ANOVA test, this graph shows a considerable amount of variation in the participants’ and assessor’s responses. In order to further analyse the reasons for this difference in response, we categorized the queries by whether they were known-item or subject queries, and by the presence of certain parts-of-speech\(^1\) in the query. The results are as follows:

- 18 percent, or about 69 of the queries marked as restrictor queries by the participants were known-item queries. Table 5.4 shows some examples of such queries:

<table>
<thead>
<tr>
<th>Examples of known-item queries that were marked as restrictor queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Integrated approach to Decision Marketing</td>
</tr>
<tr>
<td>7 Habits of Highly effective People</td>
</tr>
<tr>
<td>Assessment of Sustainable Development in Cities, vol 6</td>
</tr>
<tr>
<td>Proceedings of w060 Symposium</td>
</tr>
<tr>
<td>Asean Economic Conference</td>
</tr>
</tbody>
</table>

Table 5.4 Some known-item queries that were marked as restrictor queries

\(^1\) For a list of parts-of-speech used and their formal definitions, refer to Appendix C
The idea of a restrictor query has been defined only within the domain of subject queries, hence it does not include known-item queries. Yet, participants tended to include known-item queries as part of the domain of restrictor queries, although the difference between known-item and subject queries was explained to them during the experiment. We attribute this error to lack of understanding of the definition of a known-item query. Furthermore, participants may have been confused by the presence of certain parts-of-speech, such as prepositions and conjunctions which might have been indicators of whether a query is a restrictor query.

We also analysed the data to see how many queries were marked as restrictor queries, given the presence of certain parts-of-speech. Table 5.5 shows the results:

<table>
<thead>
<tr>
<th>Queries</th>
<th>Participants</th>
<th>Assessor</th>
</tr>
</thead>
<tbody>
<tr>
<td>with Preposition</td>
<td>RQ(%) 73</td>
<td>Not a RQ(%) 27</td>
</tr>
<tr>
<td>with Conjunction</td>
<td>42</td>
<td>58</td>
</tr>
<tr>
<td>with Adjective/Adverb</td>
<td>65</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 5.5 Comparison of responses for queries with different parts-of-speech

Based on these results, we can draw some conclusions for the variation in responses:

- Participants were generally able to identify queries with prepositions as restrictor queries, with minor discrepancies due to the selection of known item queries.
- Conjunctural queries were harder to mark as restrictor queries, as participants did not always see the relation between two terms. For instance, in the query ‘asthma’ and ‘ailment’, the assessor concluded that asthma was a kind of ailment, and thus classified the query as a restrictor query. However, this was not the case with the participants, as only one out of the five participants grading this query recognised this as a restrictor query.
- Participants often tended to over-rank queries with adjectives and adverbs as restrictor queries. This was because they failed to notice when an adjective or adverb
performed as an argument (as in the examples ‘face recognition’ and ‘quality control’) and when it performed as an adjunct (as in the examples ‘fraternal twins’ and ‘slow chemical reactions’).

5.2 Selection of core term

We present the results of the core term agreement between participants and the assessor as a contingency table that contains the following classes.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Positive (FP): Assessor fails to acknowledge a term that is classified as a core term by the participant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Negative (FN): Assessor claims that a term is a core term, but participant does not agree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Negative (TN): Both assessor and participant do not accept the term as a core term</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6 Categorization of core term agreement

For this assessment, we restrict the domain to only those queries which have been identified as restrictor queries by both participant and assessor. As each participant’s response is independent of each other, we treat each response as a separate data point. Out of a total of 625 queries, 332 queries were judged as restrictor queries by both participants and assessor. Each of these queries contained, on average 2-3 terms, which gave us a total of 844 terms. Table 5.7 gives a distribution in terms of numbers and percentages of the comparison between responses.

<table>
<thead>
<tr>
<th>Participants</th>
<th>CT</th>
<th>Not CT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>223 (26.4 %)</td>
<td>109 (12.9 %)</td>
</tr>
<tr>
<td>Not CT</td>
<td>151 (17.9 %)</td>
<td>361 (42.8 %)</td>
</tr>
</tbody>
</table>

Table 5.7 Results for core term agreement between assessor and participant
Of particular interest for our analysis are the 260 terms that fall under the categories of false negatives and false positives. Table 5.8 shows some examples of queries in which the participants did not agree with the assessor on the core term, giving the two differences.

<table>
<thead>
<tr>
<th>Query</th>
<th>Assessor’s selection of core term</th>
<th>Participants’ choice of core term(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17th century European art</td>
<td>art</td>
<td>century, European art</td>
</tr>
<tr>
<td>contemporary economies in the developing world</td>
<td>economies</td>
<td>developing world, contemporary economies</td>
</tr>
<tr>
<td>‘New Zealand’ and ‘law commission’</td>
<td>law commission</td>
<td>New Zealand</td>
</tr>
<tr>
<td>Urban studies in Singapore in the 1960s</td>
<td>Urban studies</td>
<td>studies</td>
</tr>
<tr>
<td>Economic history of Malaya</td>
<td>history</td>
<td>Malaya</td>
</tr>
<tr>
<td>20 ounces of oil</td>
<td>oil</td>
<td>ounces</td>
</tr>
<tr>
<td>Rituals in Himduism</td>
<td>hinduism</td>
<td>Rituals</td>
</tr>
</tbody>
</table>

*Table 5.8 Some queries in which there was a disagreement between participants and assessor*

We attempt to analyse why this is so, and present the analysis below:

- False negatives often occurred because participants did not always choose the term which gave the query its fundamental meaning as the core term. While the assessor’s choice was made with due consideration to generality, this was not always the case with participants’ responses. For instance, in the query ‘contemporary economies in the developing world’ the assessor chose ‘economies’ as the core term, whereas some of the participants opted for ‘developing world’. This initiated the idea that linguistic knowledge of the terms was not enough to distinguish core terms and modifiers.

- False positives occurred mainly because participants confused locative, aspectual and descriptive restrictors with core terms. Figure 5.2 gives a distribution of how many of the terms mistaken for core terms were from each of the restrictor classes.
Figure 5.2 Distribution of false-negatives according to the various restrictor classes

Locative terms were often confused with core terms in queries where the base term was not very obvious. For instance, in the query ‘Economic History of Malaya’, the assessor classified ‘history’ as the core term, and ‘Malaya’ as a locative restrictor. However, some participants classified ‘Malaya’ as the core term, and ‘Economic History’ as an aspectual restrictor.

A similar observation was made for aspectual terms. In the example ‘Rituals in Hinduism’, the assessor classified ‘hinduism’ as the core term, and ‘rituals’ as the aspectual restrictor. However, some participants were of the opinion that ‘rituals’ was the core term, and ‘hinduism’ was a descriptive restrictor.

In some cases, the participants decomposed the core term as judged by the assessor into a core term and a descriptive restrictor, simply because they were unable to decide if the descriptor functioned as an argument (and thus forms part of the core term) or an adjunct (and thus can be classified as a restrictor). This can be observed in the query ‘Urban studies in Singapore in the 1960’s’, where the assessor judged the core term as ‘urban studies’, while the participants recognized only ‘studies’ as the core term, with ‘urban’ forming a descriptive restrictor.

The observations made above for erroneous terms indicate the need for additional knowledge on the terms in order to classify them as core and restrictor. In the next chapter, we propose two ways of dealing with individual terms. The first is by using
distributional knowledge of the terms. The second is to make use of semantic knowledge of the terms such as the generality of the term in a taxonomy like WordNet. These approaches will be elaborated upon in section 6.3.

5.3 Selection of restrictors

5.3.1 Agreement between participants and assessor

For this assessment, the domain selected was only terms from queries containing the 223 true positive terms in table 5.4. A total of 414 terms were assessed. We preset the results of the classification task as a contingency table.

<table>
<thead>
<tr>
<th></th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L</td>
</tr>
<tr>
<td>L</td>
<td>106</td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
</tr>
</tbody>
</table>

Table 5.9 Contingency table to show the distribution of terms

A total of 310 queries lie on the diagonal, which indicates agreement between participants and assessor. We give explanations for some of the other observations made:

- Locative restrictors were sometimes confused for descriptive restrictors, because of their behaviour in the query. For instance, in the query ‘17th century European art’, the term European was thought to function more as a descriptor for art, rather than as denoting a location. Our proposed algorithm, thus needs to be able to differentiate locative from descriptive restrictors in such scenarios.

- Respondents faced some difficulty in judging between temporal and quantitative restrictors, sometimes confusing the two classes. This was especially evident in the case whereby the numeral consisted of four digits. For instance, in the example ‘2005 trade marks acts’, it was difficult for the participants to discern whether ‘2005’ referred to the number of trade marks acts, or to the specific year of the acts. This suggested that our proposed algorithm should be able to take care of differences in
temporal and quantitative suggestions. We propose the use of ‘named entity tagging’ to resolve this issue, as well as that faced in the previous observation.

- Often, participants confused descriptive with aspecual restrictors. Although descriptive restrictors tended to be mostly adjectives or adverbs in nature, participants faced difficulties in deciding when a term described the core term, versus when it referred to a characteristic of the core term that is not descriptive.

5.3.2 Relation between class of restrictors and linguistic behaviour

We further analysed the terms that bore agreement between participants and assessor in terms of their linguistic behaviour. We specifically looked at:

- Judging whether a particular class of restrictors were usually positioned within or outside a prepositional phrase.
- Judging whether adjectives and adverbs determined the class of restrictors.

Our findings are as follows:

- Locative and temporal restrictors tended to be positioned within the prepositional phrase (for instance, in ‘Economic history of Malaya’ and ‘European art in the eighteenth century’ respectively), while quantitative restrictors tended to be positioned outside the prepositional phrase (for instance in ’20 ounces of oil’). No specific trend was observed for descriptive and aspecual restrictors.
- Most Adjectives and adverbs either tended to be part of the core term (when they performed as arguments) or descriptive restrictors (when they performed as adjuncts).

5.4 Level of Confidence of choice

Table 5.10 gives a summary of the average confidence levels of the participants. Again, as in section 5.1, we have combined the results to reflect the confidence level as a measure of five participants alone.
Table 5.10  Summary of Levels of Confidence of the participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Standard Deviation</th>
<th>Sample Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>3.8</td>
<td>4</td>
<td>4</td>
<td>1.016001</td>
<td>1.032258</td>
</tr>
<tr>
<td>Participant 2</td>
<td>3.944</td>
<td>4</td>
<td>4</td>
<td>0.854816</td>
<td>0.73071</td>
</tr>
<tr>
<td>Participant 3</td>
<td>4.296</td>
<td>4</td>
<td>4</td>
<td>0.695887</td>
<td>0.484258</td>
</tr>
<tr>
<td>Participant 4</td>
<td>3.36</td>
<td>4</td>
<td>4</td>
<td>0.99515</td>
<td>0.990323</td>
</tr>
<tr>
<td>Participant 5</td>
<td>4.288</td>
<td>4</td>
<td>5</td>
<td>0.791242</td>
<td>0.626065</td>
</tr>
</tbody>
</table>

Figure 5.3 shows the distribution of confidence levels amongst the 25 participants.

Most participants had an average confidence level of between 3 and 5, which indicates that participants were fairly sure of the choices they made.

We also conducted an analysis of confidence levels for different types of queries, namely known item, subject queries, and those with prepositions, conjunctions and adjectives/adverbs. Figure 5.4 shows the results. Note that in this analysis, the classes are not mutually exclusive, hence a query might have been counted more than once depending on the properties it displayed.
Confidence Levels for known-item queries and conjunctional queries were relatively lower than that for other queries. Making a cross-reference to the number of known-item and conjunctional queries that were selected as restrictor queries, we might be able to infer that although users consistently tended to make the wrong decision with regards to these types of queries, they were not very confident of their answer, indicating again that their understanding of known-item and conjunctional queries was limited.
Chapter 6

Automatic Query Classification

Having given experimental evidence to show that the definitions and classification scheme proposed are replicable amongst human participants, we now propose an algorithm that distinguishes restrictor queries from other queries, and also classifies the restrictors in the query according to the proposed classification scheme.

6.1 Overview and objectives of the proposed algorithm

Figure 6.1 gives an overview of the work flow of the algorithm.

![Diagram of the proposed algorithm]

Figure 6.1 Overview of the work-flow of the proposed algorithm
The proposed algorithm has three main objectives:

- To determine whether a given query is a restrictor query;
- To identify the core term and the restrictors in a query classified as a restrictor query;
- To categorize the restrictors according to the proposed classification scheme.

We cast the above tasks as standard classification problems, which we can apply a supervised classification method. We illustrate our technique using Naïve Bayes, although any standard supervised learning algorithm can be employed with the feature set we outline later. We specifically choose to use Naïve Bayes here as it is a simple yet effective probabilistic learning technique. As identities of a query, we extract various features ranging from linguistic knowledge (such as parts-of-speech and named entity), distributional knowledge (such as call number distributions for resource titles) and semantic knowledge (generality of individual terms in a lexical taxonomy such as WordNet). We start by giving a concise description of the workings of a basic Naïve Bayes classifier.

6.2 Naïve Bayes Classifier

The Naïve Bayes classification scheme is based on Bayes’ Rule and is often used to decide a classification for an entity, given a set of observable features. Bayes’ Rule combines the prior probability of a classification and the likelihood of its occurrence to form a posterior probability, which is a conditional model of the class, given a set of features. That is:

\[
\text{Posterior Probability} = \text{Likelihood of Occurrence} \times \frac{\text{Prior probability of occurrence}}{\text{Evidence of occurrence}}
\]

Mathematically, this can be written as:

\[
P(C \mid F_1, F_2, F_n) = P(F_1, F_2, F_n \mid C) \times \frac{P(C)}{P(F_1, F_2, F_n)}
\]

where:

- C = class label from the proposed classification scheme,
- \(F_1, F_2, F_n\) = set of features that represent an instance for the classification problem.
Naïve Bayes classifiers assume ‘class conditional independence’, that is they assume that the effect of a feature on a given class is independent of the values of other features. This assumption is made to simplify the computation of the posterior probability, and is thus termed ‘Naïve’. Under this assumption, the above equation can be simplified to:

\[ P(C \mid F_1, F_2, F_n) = P(C) \times \prod_{i=1}^{n} P(F_i \mid C) \]

6.3 Features used in our algorithm

We propose to use a combination of features comprising linguistic, distributional and semantic knowledge in the classification of restrictor queries. A brief description of each of these feature sets is given below:

6.3.1 Linguistic Features

A. Parts-of-Speech

Our proposed algorithm uses parts-of-speech features of queries to deduce the probability of a given class. We specifically focus on the presence of prepositions, conjunctions, adjectives and adverbs. Although there are other parts-of-speech that may be recognizable, we choose to use only the above four rules, as we believe that these four are representative of parts-of-speech traits that are associated with and can be observed in restrictor queries. We also want the set of features proposed to be mutually exclusive, as this is a requisite for the Naïve Bayes approach.

B. Named Entity

We also propose to identify the presence of entity names, temporal expressions and numeric expressions through the process of named entity tagging.

For both parts-of-speech and named entity features, the probabilities associated with each features can be calculated from corpus counts. For instance, in order to decide whether a query is a restrictor query, we need to calculate the probability that the query is a restrictor query, given the presence of a certain feature. This can be calculated as follows:

\[ P(Q \in RQ \mid F_i) = \frac{\text{Total number of Restrictor Queries with feature } i}{\text{Total number of queries with feature } i} \]
A similar method can be used to calculate the required probabilities based on the observed feature set.

### 6.3.2 Distributional Features

Call number distributions for resource titles can provide important clues to whether a given term in a query functions as a core term or as a restrictor. Every resource in a library, be it a book or a multimedia article, has a call number that is unique to that resource. Call numbers not only help users to locate a specific resource, but are also useful to librarians for cataloguing purposes. Our hypothesis is that the more scattered the call number distribution for a given term is, the less likely that the term forms a distinct subject in the library, and thus the less likely that it is a core term. We thus propose to use the call number distribution in calculating the probability that a given term is a restrictor.

Call number classification schemes are based on various combinations of the resource’s subject area, author name and publication year, and often map to a unique identification number. Modern classification systems are hierarchically enumerative, so that successive characters in the call number of a resource describe it in increasing levels of specificity. Most libraries use one of two classification standards: the Dewey Decimal Classification system (DDC), and the Library of Congress Classification (LCC) system. The LCC is adopted primarily by academic research libraries, and is used by the LINC system.

The LCC presents call numbers as alphanumeric sequences that describe the subject area and the author of the publication. The typical LCC notation contains a mixed notation of one to three letters, followed by one to four integers, and possibly a short decimal fraction, up to three digits in length. An example of such a notation is: `ASG567.34`

Library call numbers serve two functions. The classification notation groups related materials together. The second portion of the complete call number uniquely identifies different works in the same class. While LLC adopts a ‘Cutter Number’ system devised by Cutter (1962), which uniquely identifies an author name by his initial letter, and a number sequence comprising two digits, LINC uses a slightly different method, which
gives the first three letters of the author’s name, followed by the year of publication of the resource. Thus, a typical call number in LINC would look as shown in Figure 6.2:

![Call Numbers as seen in LINC.](image)

Since the second part of the call number refers to the author name and year of publication, this part is unique to each resource. As our objective in analysing the call number distribution is to see whether a certain term gives a condensed or scattered distribution, we focus only on the first, topical part of the call number and ignore the second, author-specific part.

We propose to first use a multiplicative method to convert the entire sequence of alphanumeric characters into a string of numerals. This is done in order to judge the proximity between two resources in the next step. Once all the relayed call numbers have been changed to a numeric format, the spread of the distribution can be observed by calculating its standard deviation. The higher the standard deviation, the more dispersed is the distribution, thus the lower is the probability that the term is a restrictor. Conversely, the lower the standard deviation, the more condensed is the distribution, thus the higher is the probability of the term being a restrictor.

### 6.3.3 Semantic Features

We also propose to use semantic features of a term such as how generic one term is as compared to another in order to decide the particular role that a term plays in a query. One approach for computing the generality factor is to compare terms in a taxonomical structure such as WordNet. Many algorithms have been proposed to derive the semantic similarity between two terms in the WordNet taxonomy (Resnik (1995); Jiang, and Conrath (1997); Hirst and St-Onge (1998); Leacock and Chodorow (1998)). Of particular
interest is Resnik (1995)’s work which uses the concept that ‘the semantic distance between a pair of concepts can be judged by the extent to which they share information’. He defined the similarity between two concepts in WordNet to be the information content (IC) of their lowest super-ordinate (lso).

Mathematically,
\[
\text{sim}(c_1; c_2) = -\log p(lso(c_1; c_2))
\]

where \( p(c) \) = Probability of encountering an instance of a concept \( c \) in some specific corpus.

He also hypothesized that the closer a node was to the root of the taxonomy, the lower was its information content. This idea forms the basis for our determination of a core term and restrictor, as the value of the information content of a node gives an idea of the relative distance between that node and the root, thereby indicating the generality of the term (terms with a lower information content value will be more generic than terms with a higher information content value.)

Mathematically, the information content value bears a direct relation to the probability that a term is a restrictor, and can be computed as follows:

\[
P(X: \text{restrictor} \mid \text{relative position of } X \text{ in Wordnet}) = \frac{\text{IC of node } X}{\text{IC of lower-most node in Wordnet}}
\]
6.4 Execution of Algorithm

6.4.1 Query Classification

Figure 6.3 gives a detailed flow of the process that takes place during query classification.

![Flowchart of Query Classification](image)

Figure 6.3: Flow-chart to show Query Classification at Runtime

In order to decide whether a query is a restrictor query, we calculate the probability that it is a restrictor query, given a set of observable features, and compare this to the probability that it is not a restrictor query. These probabilities are calculated using corpus counts, and by the Naïve Bayes rule as described earlier. The query is first passed through a part-of-speech tagger, and then through a named entity tagger. These two processes help determine the linguistic features characteristic of this query. The probability that the query is a restrictor query, given these specific features is then
calculated using Naïve Bayes rule. (values for the RHS of the Bayes rule come from corpus counts). A similar process is carried out to determine the probability that the query is not a restrictor query, given these same observed features. The two probabilities are finally compared to decide whether the given query is a restrictor query or not.

### 6.4.2 Restrictor Identification

Once a query has been classified as a restrictor query, the second phase of the algorithm determines the core term and the restrictor. For simplicity, we will assume here that our example consists of only two terms, X and Y, of which one will be deemed the core term and the other the restrictor at the end of the process. From an assessment of a random sample of 300 restrictor queries in LINC, it was observed that close to 75 percent of the queries comprised more than one restrictor. Hence it is necessary to scale our approach to compare multiple terms. This will be addressed in the next section.

Figure 6.4 outlines the process of determining the base term and the restrictor
The terms X and Y are assessed in parallel to determine a probability that X is the restrictor, and the probability that Y is a restrictor. Besides using the linguistics features to determine this probability, we also make use of the distributional and semantic features proposed in the previous section. The final probability that X or Y is a restrictor is calculated as a product of these three distinct features. The two probabilities are then compared to determine which of the terms is a restrictor.

### 6.4.3 Restrictor Classification

Figure 6.5 shows the procedure that takes place during the classification of a restrictor

![Flow-chart showing the procedure of Restrictor classification at runtime](image)

Again, using Naïve Bayes for the linguistic features discussed in the previous section, the probability that the restrictor R belongs to a certain class of restrictors is calculated for each of the classes: Locative, Temporal, Quantitative, Descriptive, and Aspectual. The
probabilities are then compared to determine which of them is the biggest, which determines which class of restrictors $R$ belongs to.

### 6.5 Scaling the Algorithm

As mentioned earlier, often queries contain more than one core term and restrictor, hence it is necessary for our algorithm to be able to handle comparisons of multiple terms. We present an exhaustive search approach for doing so, with its advantages and disadvantages.

<table>
<thead>
<tr>
<th><strong>Method</strong></th>
<th><strong>Exhaustive Search</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Identify all terms in a query.</strong></td>
<td><strong>For each term, compare the term to every other term in the query.</strong></td>
</tr>
<tr>
<td><strong>At the end of the iteration, retain the term that was classified as the base term.</strong></td>
<td><strong>After all comparisons, the term that was classified as the base term the most number of times will be the base term.</strong></td>
</tr>
</tbody>
</table>

**Example**

*(Economic History of Singapore)*

- **Iter 1**: compare *history* and *economic*. *Economic* is the base. Retain *economic*.
- **Iter 2**: compare *history* and *Singapore*. *History* is the base. Retain *history*.
- **Iter 3**: compare *economic* and *Singapore*. *Singapore* is the base. Retain *Singapore*
- Compare the number of times each term has been classified as a base term. As *History* has been classified the most number of times, it is chosen as the base term.

<table>
<thead>
<tr>
<th><strong>Time Complexity</strong></th>
<th>$O(n^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages and Disadvantages</strong></td>
<td>As there is a comparison between all the terms in the query, the final core term selected will be the most generic term of all the terms.</td>
</tr>
<tr>
<td></td>
<td>Expensive, as the efficiency of the algorithm is of the order $O(n^2)$. However, given that the number of terms to be compared is small (on average 3-4 terms per query), the efficiency is not greatly affected for small $n$.</td>
</tr>
</tbody>
</table>

*Table 6.1 Description of exhaustive search method to scale the proposed algorithm*
6.6 A real life example

Figures 6.6 to 6.____ show the expected analysis of a real life example, ‘European Literature in the eighteenth century’. For further examples on the expected results for other queries, please refer to Appendix C.

![Diagram of query identification process](image-url)

**Figure 6.6 Query identification**
Figure 6.7 Three iterations to compare the terms of the query, after which literature is determined as the core term.
Calculate $P(R|Ci | Adj, Loc)$

**Start**

European: $P(R|Loc | Adj, Loc)$

Compare probabilities

Is $P(R|Loc | Adj, Loc)$ the highest prob?

<Yes>

European is a Locative restrictor

Finish

Eighteenth century: $P(R|Tem | Prep, Tem)$

Compare probabilities

Is $P(R|Tem | Prep, Tem)$ the highest prob?

<Yes>

Eighteenth century is a temporal restrictor

Finish

Figure 6.8 Classification of the restrictors into the desired classes
Chapter 7

Conclusion

We conclude our discussion of restrictor queries after highlighting some limitations of our research and areas for future study.

7.1 Limitations of our study

In this study, we only take into account subject queries keyed into the system via the simple keyword search. Although most restrictor queries can be obtained from this domain, yet another way in which queries can be limited is when the catalogue user makes use of multiple known fields (for instance, the author and title fields together). Thus, in a sense, the domain of restrictor queries has not been explored to its full potential.

A major drawback of this study is that although we have proposed an algorithm for automatically classifying queries, we have no means to validate our assumptions, as the algorithm has not been implemented. Yet, what we would like to focus on here is the importance of the feature set used to classify queries and restrictors. The Naïve Bayes theorem is simply a means of implementing this feature set, and need not necessarily be the best method of doing so. We welcome suggestions on more effective classification algorithms for restrictor queries.
7.2 Areas for future research

In this study, we have mostly considered only features specific to the query itself (such as linguistic and semantic knowledge). We also gave a sneak preview of how the library structure might be used in our assessment of restrictor queries by examining call number distributions. In reality, the way in which a library is structured plays a huge role in the determination of a restrictor query. Thus, we can look at features such as classification of titles, subject areas and other forms of metadata to refine our automatic classification of restrictor queries.

Exploring the domain of known field searches can also form a way in which the notion of a restrictor query can be expanded.

From a broader perspective, we have only proposed one way of dealing with subject queries. The domain of subject queries itself is very vast and restrictor queries do not take into account all of these queries. Thus, through our research, we hope that we have opened up a new field for thought in OPAC query classification.

7.3 Conclusion

Realizing that there is a problem articulating the user’s intentions when he enters a subject query into an OPAC, we propose the notion of a restrictor query as a means of determining the semantics of a query. We give a formal definition for a restrictor query, as well as a comprehensive classification scheme for restrictors. We then evaluate these definitions against human participants to test if they are replicable. The results showed us that participants generally agreed with each other in their responses, and were quite confident of their choices. Finally, we propose an algorithm for automatic classification of restrictor queries based on the Naïve Bayes theorem.

Overall, it would be prudent to say that the research conducted in this thesis has been a successful endeavour and is a step forward in the right direction towards building a comprehensive system that enriches the online search experience for OPAC users.
References


Appendices

Appendix A
Glossary of Terms

**Adjective:** A word that modifies a noun or a pronoun by describing, identifying, or quantifying words.

**Adjunct:** An optional, secondary element in a sentence construction which, when removed, does not affect the structure identity of the rest of the sentence.

**Adverb:** A word that modifies a verb, an adjective, another adverb, a phrase, or a clause

**ANOVA:** Acronym for Analysis of Variance, which is a test of the statistical significance of the differences among the mean scores of two or more groups on one or more variables

**Argument:** A primary entity in a sentence construction that bears a direct semantic relation to the head term of the sentence

**Conjunction:** A word that link words, phrases, and clauses in a sentence.

**F-Value:** Mathematical value used in an ANOVA test to determine the variation in the average scores of two or more groups.

**Information Content:** A term coined by Resnik (1995) which computed the value of the semantic distance between a pair of concepts by the extent to which they share information

**Interjection:** A word added to a sentence to convey emotion. It is not grammatically related to any other part of the sentence.

**Library of Congress Classification Scheme:** A popular resource classification scheme used in many academic libraries that presents call numbers as alphanumeric sequences that describe the subject area and the author of the publication.

**LINC:** Acronym for Library Integrated Catalogue, which is the online portal used to maintain information on library resources at the National University of Singapore.

**Named-Entity Tagging:** A process by which words denoting dates, time, locations and quantities are tagged as such.

**Noun:** A word used to name a person, animal, place, thing, and abstract idea
**OPAC**: Acronym for Online Public Access Catalogue, which refers to the electronic version of library catalogue.

**Preposition**: A word that links nouns, pronouns and phrases to other words in a sentence.

**Pronoun**: A word that can replace a noun or another pronoun.

**P-Value**: the probability in an ANOVA test that the variation between groups obtained is due to random factors.

**Restrictor query**: A subject query which consists of at least one core term (comprising the argument and its associated head word or phrase) and at least one adjunct which limits the scope of the core term.

**Verb**: A word which asserts something about the subject of the sentence and express actions, events, or states of being.

**WordNet**: A lexical reference system that contains nouns, verbs, adjectives and adverbs that are grouped together into sets of cognitive synonyms, each of which expresses a distinct concept.
Appendix B
Sample questionnaire handed out to participants

1. ‘kinship’ and ‘marriage’

Is this a Restrictor Query? Yes No

If Yes, the base term is: ___________________________

Indicate the restrictors and the type below:

<table>
<thead>
<tr>
<th>Restrictor</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How sure are you of your choice?

☐ Not sure at all ☐ Not very sure ☐ Neutral ☐ Sure ☐ Very sure

2. Biology of Marine Species

Is this a Restrictor Query? Yes No

If Yes, the base term is: ___________________________

Indicate the restrictors and the type below:

<table>
<thead>
<tr>
<th>Restrictor</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How sure are you of your choice?

☐ Not sure at all ☐ Not very sure ☐ Neutral ☐ Sure ☐ Very sure

3. Asset allocation

Is this a Restrictor Query? Yes No

If Yes, the base term is: ___________________________

Indicate the restrictors and the type below:

<table>
<thead>
<tr>
<th>Restrictor</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How sure are you of your choice?

☐ Not sure at all ☐ Not very sure ☐ Neutral ☐ Sure ☐ Very sure
Appendix C
Examples of automatic query classification

In this appendix, we detail the flow of our algorithm by using some real life examples. At each stage, we also give the expected result.

**Example 1: Material properties of silicon**

**Procedure:**

- **Query Identification**
  - Run query through POS and NE taggers
  - Determine linguistic feature set for the query, in this case it is *Adj* and *Prep*
  - Calculate
    \[
    P(Q \in RQ \mid Q \subset Adj, Prep \text{ and } Q \not\subset Adv, Conj, date, time, location, quantity) \quad & \quad P(Q \not\in RQ \mid Q \subset Adj, Prep \text{ and } Q \not\subset Adv, Conj, date, time, location, quantity)
    \]
  - Compare the probabilities
  - Is \( P(Q \in RQ) > P(Q \not\in RQ) \rightarrow Yes \)
  - Query is a restrictor query

- **Restrictor Identification**
  - For term ‘Material properties’(X):
    - Determine linguistic feature set for the term, in this case it is *Adj*
    - Calculate \( P1 = P(X:res \mid Adj, \text{ all other properties absent}), P2 = P(X:res\text{call No. dist of X}) \) And 
      \( P3 = P(X:res\text{ generality of X in WordNet}) \)
    - Calculate \( P(X: res) = P1 \times P2 \times P3 \)
  - For term ‘silicon’(Y):
    - Determine linguistic feature set for the term, in this case it is *Noun*
    - Calculate \( P1 = P(Y:res \mid Noun, \text{ all other properties absent}), P2 = P(Y:res\text{call No. dist of Y}) \) And 
      \( P3 = P(Y:res\text{ generality of Y in WordNet}) \)
    - Calculate \( P(Y: res) = P1 \times P2 \times P3 \)
  - Compare both probabilities. Is \( P(X: res) > P(Y: res) \)? \( \rightarrow \) Material Properties os the restrictor and silicon is the core term

- **Restrictor Classification**
  - For the term ‘Material properties’
    - Calculate \( P(Y:res \text{ of class Ci} \mid Adj, \text{ all other properties absent}), \) where Ci can be Locative, temporal, quantitative, descriptive or aspectual
    - Compare the probabilities. Is \( P(Y:\text{aspectual restrictor} \mid Adj, \text{ all other properties absent}) \) the highest? \( \rightarrow Yes \)
    - Material properties is an aspectual restrictor
Example 2: Festivals in Thailand

Procedure:

- **Query Identification**
  - Run query through POS and NE taggers
  - Determine linguistic feature set for the query, in this case it is *Adj, Prep* and *Loc*
  - Calculate
    \[ P(Q \in RQ | Q \subset Adj, Prep, Loc \text{ and } Q \notin Adv, Conj, date, time, quantity ) & \]
    \[ P(Q \notin RQ | Q \subset Adj, Prep, Loc \text{ and } Q \notin Adv, Conj, date, time, quantity) \]
  - Compare the probabilities
  - Is \( P(Q \in RQ) > P(Q \notin RQ) \) → Yes
  - Query is a restrictor query

- **Restrictor Identification**
  - For term ‘Festivals’ (X):
    - Determine linguistic feature set for the term, in this case it is *Noun*
    - Calculate \( P(X:res | Noun, \text{ all other properties absent}) \), \( P2 = P(X:res| \text{ all other properties absent}) \)
      - \( P3 = P(X:res| \text{ generality of X in WordNet}) \)
    - Calculate \( P(X:res) = P1 \times P2 \times P3 \)

  - For term ‘Thailand’ (Y):
    - Determine linguistic feature set for the term, in this case it is *Noun, Loc*
    - Calculate \( P(Y:res | Loc, Noun, \text{ all other properties absent}) \), \( P2 = P(Y:res| \text{ all other properties absent}) \)
      - \( P3 = P(Y:res| \text{ generality of Y in WordNet}) \)
    - Calculate \( P(Y:res) = P1 \times P2 \times P3 \)

  - Compare both probabilities. Is \( P(X:res) > P(Y:res) \)? → No
  - Festivals is the core term and Thailand is the restrictor

- **Restrictor Classification**
  - For the term ‘Thailand’
    - Calculate \( P(Y:res \text{ of class } Ci | Noun, Loc, \text{ all other properties absent}) \), where
      - \( Ci \) can be Locative, temporal, quantitative, descriptive or aspectual
    - Compare the probabilities. Is \( P(Y:\text{locative restrictor} | Noun, Loc, \text{ all other properties absent}) \) the highest? → Yes
    - Thailand is a locative restrictor
Example 2: A pound of flesh

Procedure:

• Query Identification
  - Run query through POS and NE taggers
  - Determine linguistic feature set for the query, in this case it is Prep and Quan
  - Calculate
    \[ P(Q \in RQ \mid Q \subset \text{Prep, Quan and } Q \not\subset \text{Adv, Adj, Conj, date, time, location} ) \& P(Q \not\in RQ \mid Q \subset \text{Prep, Quan and } Q \not\subset \text{Adv, Adj, Conj, date, time, location} ) \]
  - Compare the probabilities
  - Is \( P(Q \in RQ ) > P(Q \not\in RQ) \) → Yes
  - Query is a restrictor query

• Restrictor Identification
  For term ‘pound’(X):
  - Determine linguistic feature set for the term, in this case it is Noun, Quan
  - Calculate \( P1 = P(X:\text{res} \mid \text{Noun, Quan, all other properties absent}), P2 = P(X:\text{res| call No. dist of X}) \) And
    \( P3 = P(X:\text{res| generality of X in WordNet}) \)
  - Calculate \( P(X: \text{res}) = P1 \times P2 \times P3 \)

  For term ‘flesh’(Y):
  - Determine linguistic feature set for the term, in this case it is Noun
  - Calculate \( P1 = P(Y:\text{res} \mid \text{Noun, all other properties absent}), P2 = P(Y:\text{res| call No. dist of Y}) \) And
    \( P3 = P(Y:\text{res| generality of Y in WordNet}) \)
  - Calculate \( P(Y: \text{res}) = P1 \times P2 \times P3 \)

  - Compare both probabilities. Is \( P(X: \text{res} ) > P(Y: \text{res}) \)? → Yes
  - Flesh is the core term and pound is the restrictor

• Restrictor Classification
  For the term ‘pound’
  - Calculate \( P(Y:\text{res of class } Ci \mid \text{Noun, Quan, all other properties absent}), \) where
    \( Ci \) can be Locative, temporal, quantitative, descriptive or aspectual
  - Compare the probabilities. Is \( P(Y: \text{quantitative restrictor } \mid \text{Noun, quan, all other properties absent}) \) the highest? → Yes
  - Pound is a quantitative restrictor