Chapter 1
Introduction

1 From Archie to Google and Beyond

Information retrieval is a dynamic discipline with a rich history. However, for much of its history, it had little or no impact on people's everyday lives. Many of the earliest consumers of information retrieval technologies were government researchers, scientists, and librarians. That began to change after the invention of the World Wide Web in the early 1990s.

Before the introduction of the Web, a number of information sources were available online. Most of the online information was published and controlled by government organizations or academic institutions. It was uncommon for everyday citizens to use these online systems, let alone publish their own content. The Web revolutionized the way that information was published. It allowed individuals and organizations to create content that was instantly available and easy to access. It also provided a way of linking content together, which was not possible with the older online systems. As computing costs decreased and online popularity increased, the amount of information available on the Web exploded.

As more electronic documents started appearing online, a natural desire to search the content arose. Various search tools were developed to help users find relevant files and documents. The earliest Internet search tools, Archie, Gopher, Veronica, and Jughead allowed users to search FTP servers. However, the popularity of FTP waned after the introduction of the Web. This ushered in a new era that gave rise to Web search engines. Unlike their predecessors, which were used by small fractions of the population, Web search engines such as Google are used every day by millions of users across the globe. Therefore, what started as a small, relatively unknown field of study, has evolved into an integral part of modern society.

2 The Academic and Industrial Perspectives

Yahoo and Google were both grown out of academic research projects. They currently are the two most popular commercial Web search engines in the United States.

D. Metzler, *A Feature-Centric View of Information Retrieval*,
The Information Retrieval Series 27,
DOI 10.1007/978-3-642-22898-8_1,

Clearly, the academic research community, in the early days of the Web, was developing cutting edge search technologies. However, as the commercial search engines came of age, it become increasingly difficult for the academic researchers to keep up with the collection sizes and other critical research issues related to Web search. This caused a divide to form between the information retrieval research being done within academia and industry.

There are several reasons for this divide. First, as commercial search engines mature, they are able to collect more data in the form of query logs, click-through patterns, and other types of user data which are invaluable to Web search. The companies have little incentive to release these data to academia, especially amid growing privacy concerns. Second, commercial search engines have much more computing power than most academic research institutions. Therefore, they are able to crawl more Web pages, build larger indices, use real data streams, and experiment with much more costly computations. Finally, commercial search engines are very protective of their search algorithms and techniques and do not typically publish their findings in scholarly conferences and journals. This is not surprising, since revealing technical details of ranking functions may allow spammers and other malicious entities to adversely influence search results.

To put things into perspective, let us compare academic and industrial collection sizes. The Text REtrieval Conference (TREC), which was started in 1992, provides a set of standard, reusable test collections (i.e., document collection, queries, and relevance judgments) that most academic information retrieval researchers use when evaluating retrieval systems. One of the largest TREC test collections, called GOV2, is a 2004 crawl of the .gov top level domain. It consists of approximately 25 million Web pages (428 GB of text). In comparison, it is believed that Google has upwards of 25 billion items in its index, which is 1,000 times larger than GOV2. In an attempt to reduce the academic-industrial gap, in terms of collection sizes, the ClueWeb09 collection set was recently released, which consists of approximately 1 billion Web pages.

One of the goals of this work is to reduce the divide in understanding that exists between academic and commercial information retrieval systems with respect to large data sets. Many of the techniques and ideas developed here have been inspired by large test collections, such as GOV2 and ClueWeb09. While these collections are admittedly not Web-scale, they are significant and sizable improvements over the test collections that have been used to develop most of the current state-of-the-art information retrieval models.

3 Paradigm Shifts

As we just alluded to, large collections, such as those handled by commercial search engines, provide a new set of challenges for information retrieval researchers. In this work, we describe highly effective information retrieval models for both smaller, classical data sets, and larger Web collections. As we will show throughout this

work, the current state-of-the-art academic retrieval models are not robust enough to achieve consistently effective retrieval results on large collections.

Most of the these models are based on the so-called "bag of words" assumption. Under this assumption, text (e.g., queries and documents) are represented as unordered sets of terms. This means that any notion of term ordering is lost. For example, under this representation, the texts *the bear ate the human* and *the human ate the bear* are identical. However, these pieces of text clearly have different meanings. While this is an overly simplistic representation, very few have been able to develop non-bag of words retrieval models that are consistently and significantly better than the state-of-the-art bag of words models. Many researchers over the past few decades have tried in vain, but there has been very little success.

The ranking functions associated with bag of words retrieval models often consist of some combination of *term frequency* (TF) and *inverse document frequency* (IDF). The IDF component acts to discriminate between informative and non-informative query terms. Those terms that have a high IDF are considered more informative, because they rarely occur in the collection. On the other hand, terms that have a low IDF are considered uninformative, since they occur in many documents. As the number of documents in a collection increases, IDF becomes increasingly important in order to discriminate between those documents that only contain non-informative query terms and those that contain highly informative query terms.

On the other hand, the TF component, which is often normalized in some way with respect to the *document length*, is used to discriminate between documents that contain a query term several times and those that contain the term many times. This makes the assumption that documents that contain more mentions of a given query term are more "about" the given term and therefore are more likely to be relevant to the query. As we will discuss shortly, this is a bad assumption, especially as collection sizes increase and documents become noisier. The TF component becomes more important as documents get longer, since query terms are unlikely to occur more than one time in a very short document, and since long documents are more likely to contain more diverse term occurrence statistics.

Therefore, the TF and IDF components used within bag of words ranking functions, when combined together, discriminate along two dimensions—informativeness (IDF) and aboutness (TF). However, when dealing with large Web collections, a third dimension that we call noisiness enters the picture.

All collections, even small ones that consist entirely of news articles, contain some noise. However, large Web collections are likely to contain abundant amounts of noise. The standard TF and IDF features are not enough to overcome this noise. In fact, these features may actually help amplify the noise in some cases. Let us consider the query *habitat for humanity* run against a large Web collection. Using a state-of-the-art bag of words retrieval model, many of the top ranked results are relevant to the request. However, there are several results very high in the ranked list that do not contain a single occurrence of the term *humanity*. Instead, these documents contain hundreds of occurrences of the high IDF term *habitat*. These documents are ranked so highly because they contain many occurrences of a very high IDF term.

Documents that contain hundreds of occurrences of some high IDF term are going to result in poor, noisy matches for most bag of words models based on TF and IDF. Such documents may arise by coincidence, or a spammer who wishes to increase the ranking of a given Web page may "stuff" the page with such terms. In either case, it is very undesirable for these documents to be ranked highly.

Another more subtle way that noise may be introduced into bag of words matches happens when two or more query terms match a document, but the matches are random or unrelated to the query. For example, in the *habitat for humanity* case, consider a document that contains a paragraph that discusses habitat changes caused by global warming and another paragraph that discusses the negative impacts of global warming on humanity. Both the terms *habitat* and *humanity* will match this document, but the matches are unrelated to the query. That is, the terms just happened to match by chance. This is another example of noisy matches that can arise in large collections. In fact, as collection size grows, so does the chance that any two query terms will randomly match within some document.

Hence, new ranking function components, above and beyond TF and IDF must be used in order to reduce the number of noisy matches. There are a few ways to address this issue. First, one of the simplest ideas is to cap the TF component and not allow it to grow unbounded. While this addresses some noise issues, it fails to address the problem of randomly matching query terms. Second, in order to address the so-called term stuffing problem, anti-spam techniques may be developed in order to automatically detect malicious or misleading content. However, like capping TF, this only addresses some of the noise issues. Finally, term proximity features may be used in order to ensure that matches are not random and that they are somehow related to the query. For example, these types of features could be used to promote documents that contain the exact phrase *habitat for humanity* as opposed to those that simply contain random occurrences of the terms *habitat* and *humanity* on their own. It is this third option that we heavily explore within this work in order to overcome the limitations imposed by TF and IDF alone. It is important to notice that by using term position information, we are abandoning the bag of words assumption and move to a richer, more realistic text representation.

Aboutness, informativeness, and noisiness reflect the three primary information retrieval paradigm shifts. Here, a paradigm shift is a new way of approaching a problem with a given set of characteristics. The paradigm shifts are summarized in Fig. 1.1. The figure plots three data sets (CACM, TREC Disks 1 and 2, and GOV2) with respect to their average document length and the number of documents in the collection. As the figure shows, the TF paradigm shift moves along the average document length axis and the IDF shift moves along the number of documents axis. We also see that the noise shift moves along both axes, but is only present for large collections, such as GOV2. The newer ClueWeb09 collection, which is not shown in this plot, will likely require another (yet to be discovered) paradigm shift in retrieval methodologies to achieve maximal effectiveness.

We hypothesize that many of the previous attempts to go beyond the bag of words assumption have failed because of the small data sets used. In fact, most, if not all, of the previous research on non-bag of words model have been evaluated on test

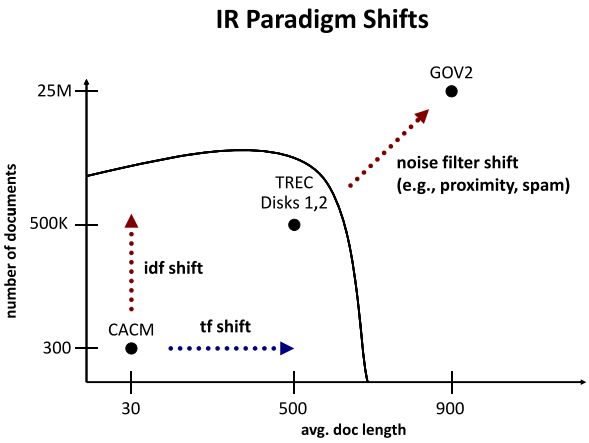

Fig. 1.1 A summary of the three primary information retrieval paradigm shifts. They include the TF shift (aboutness), the IDF shift (informativeness), and the noise shift (noisiness)

collections within the region shown in Fig. 1.1. Poor, or inconclusive results, were achieved because the data sets did not exhibit the characteristics necessary to exploit the noise reducing features associated with non-bag of words models. Therefore, new models that go beyond the bag of words assumption should be tested on large, noisy data sets in order to properly evaluate their full potential.

4 A Robust Retrieval Model

In this work, we describe a robust statistical information retrieval model based on Markov random fields. In particular, the model is designed to support the following desiderata:

1. Support basic information retrieval tasks (e.g., ranking, query expansion, etc.).
2. Easily and intuitively model query term dependencies.
3. Handle arbitrary textual and non-textual features.
4. Consistently and significantly improve effectiveness over bag of words models across a wide range of tasks and data sets.

The model we describe goes beyond the bag of words assumption in two ways. First, the model can easily exploit various types of dependencies that exist between query terms. This eliminates the term independence assumption that often accompanies bag of words models. Second, arbitrary textual or non-textual features can be used within the model. Thus, it is possible to use simple features such as TF and IDF, or more complex features, such as those based on term proximity. Other possible features include PageRank, inlink count, readability, spam probability, among others. None of the current state-of-the-art models allow arbitrary features to be incorporated as easily as the Markov random field model.

As we will show, combining term dependencies and arbitrary features results in a very robust, powerful retrieval model. Within the model, we describe several extensions, such as an automatic feature selection algorithm and a query expansion

framework. The resulting model and extensions provide a flexible framework for highly effective retrieval across a wide range of tasks and data sets.

5 Outline

The remainder of this work is laid out as follows.

- Chapter 2 summarizes several important classical retrieval models. The chapter divides the discussion of the models into non-bag of words models (e.g., Binary Independence Model, BM25, language modeling, etc.) and models that go beyond the bag of words assumption (n-gram language models, Inference Network Model, etc.). It concludes with a brief discussion of the current state-of-the-art.
- Chapter 3 motivates and covers the basics of the Markov random field model for information retrieval, which serves as the basis for exploring feature-based retrieval models throughout the remainder of the work. The chapter begins with a theoretical treatment of the model and concludes with a detailed analysis of the practical aspects of the model.
- Chapter 4 explains how the Markov random field model can be used for robust, highly effective feature-based query expansion via the use of a method called Latent Concept Expansion (LCE). LCE and several of its extensions are described in detail.
- Chapter 5 describes a powerful extension of the basic Markov random field model that supports query-dependent term weighting. Unlike most existing retrieval models that weight all features the same, regardless of the query, the approach described in this chapter provides a means for adaptively weighting the importance of each feature based on properties of the query.
- Chapter 6 covers a number of techniques that can be used to estimate the parameters of feature-based models. The emphasis of this chapter is on simple techniques that can be used to learn the parameters of models typically encountered and used within research environments. More sophisticated approaches that are more well-suited for estimating parameters within industrial settings are also covered.

In addition to the primary technical content, this work also includes the two following appendices that are meant to provide readers with additional background information.

- Appendix A covers the anatomy of TREC test collections and summarizes the data sets used throughout this work.
- Appendix B explains how the various data sets used throughout this work are computed.