Empirical Analysis of Diversity in the Web

Suneel Kumar Kingrani

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Computer Science and Information Systems
Birkbeck, University of London
United Kingdom

DECLARATION

This thesis is the result of my own work, except where explicitly acknowledged in the text. _____ [Suneel Kumar Kingrani].

ABSTRACT

In the pre-Internet age, radio, newspapers and television were the only information sources available to people. Currently, over the Internet, the sources of information are not limited to these few channels but have grown exponentially. To fulfil the information appetite of any user, the Internet has a vast and diverse set of information sources available. Having considered the exponential growth of diverse sets of information over the Internet, it is natural to consider methods to evaluate or measure the diversity of information for any topic, specifically when the topic or query for the required information is ambiguous or vaguely specified.

Ambiguous queries, in Web search engines, hide the actual information needs of users. When the actual information needs of users are uncertain then it is preferable to present the users with a diverse and less redundant set of search query results. There arises a cyclic process of abandoning and retyping a search query when users do not find the required information in search query results. The diversity of search query results makes it less likely that a user will abandon a query in a Web search engine.

In this thesis we provided a new perspective to analyse and examine the diversity into the Web, which is not done previously. To do that we introduce the methods of diversity, namely *Inverse Simpson's index* and *Shannon's diversity index*, which includes *Richness* and *Evenness*, into the context of the Web. After that we investigate the relationship between diversity related factors and the prediction of lifetime of a query popularity, with the help of Cox proportional hazard regression model. Afterwards we propose an alternate method to calculate the diversity for overlapping categories. Along with that we further analyse the trend and seasonality for any changes in diversity with respect to time. Finally we apply these diversity methods to two application domains: (i) to get the optimal number of clusters for text data and (ii) to produce a meta-evaluation of evaluation methods for diversified search.

In summary, we empirically analyse diversity and its usefulness over the Web. This can help understand the Web and get more useful and relevant information from the huge data available over the Web.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

This thesis is dedicated to my parents.

ACKNOWLEDGEMENTS

First and foremost I express my sincere gratitude to my supervisors, Prof. Mark Levene and Dr. Dell Zhang, for their continuous support and guidance. Without them, this thesis would not have been possible. It is their experience, patience and technical as well as personal guidance that made this thesis into a quality work. I would like to extend my thanks to Prof. George Loizou, who has been very generous in his guidance towards my thesis structure, content and proofreading and his insightful comments and encouragement made me focus in the right direction.

I thank all my teachers and mentors throughout my academic career. It is all because of them that made me what I am today.

I shall thank my mother and father as their emotional, moral and spiritual candle always light up the way in my life. I would like to acknowledge the support of my whole family who has been incredibly supportive and understanding throughout my life's journey.

I am grateful to all my friends for always being there. Particularly, I thank my labmates who were always in for the stimulating and interesting discussion on all the topics in the world.

CHAPTER 1

INTRODUCTION

1.1 PERSPECTIVE OF DIVERSITY OVER THE WEB

After the invent of the Internet, there has been an exponential increase in the multiple channels, i.e. websites, which provide information about any query posted by different users over the Internet.

Web search engines are nowadays the primary means for people to locate required information from the vast and diverse set of sources or websites available over the Internet [49]. A survey conducted in 2012 reported that about 91% of users had started to employ search engines for their information needs [65]. N. Craswell et al. [25] observed that a user's click on search results is biased towards top search results which means that a website effectively does not exist if it is not included in the top search results.

Due to different backgrounds of users, the queries posted by them are ambiguous, i.e. queries that are not clear to understand and having more than one meaning. For example, a user posts a query "Donald Trump". Now due to no knowledge of the user's intentions, it is not clear for example, if the user is interested in "the history and background of Donald Trump", "the election campaign of president Donald Trump" or "a particular policy statement from Donald Trump, e.g. the ban on some of the mostly Muslim countries".

To show more relevant results for ambiguous queries, Web search engines use the personalisation of search results [63], in which the results are personalised to every user based on there personal information, previous search history and the preference of the websites that a user clicks in the search

results. There is a well-known problem with personalization of search results which is termed as filter bubble [61]. It is the state in which a user is presented with the only results which coincide with a user's personal opinions and forces one belief over another belief which might go against a user's personal opinion. In this way, the personalization causes an intellectual isolation of a user from a diverse range of information available on any topic over the Internet. There is also an issue of privacy which is linked with gathering personal information of any user. For example, When AOL publicly released their anonymous search terms for the purpose of research. For which they were sued by AOL subscribers [55] who found that search queries from that anonymous dataset links back to them.

In order to cover multiple aspects of ambiguous queries, it is important to present the users with a diverse set of information sources. Having different users' information needs that are attached to similar queries is generally insufficient to present results based only on relevance to the query in question [93].

This gives rise to a crucial question of whether the top K results provided by a search engine are relevant to different user's needs and at the same time contain the diverse sources of information over the web.

In recent years, there has been a surge of research on diversifying Web search results [28, 74]. Carbonell and Goldstein [14] were the first to recognise this problem, and they proposed a method which focuses on the similarity between documents in a result set, to introduce novelty, along with the relevance of these documents to the query. By considering novelty in the resulting documents, they introduce diversity in the result set, so as to satisfy different users' information needs attached to the query.

Clarke et al. [21] proposed the extended version of the nDCG algorithm for evaluation of the list of ranked documents, i.e. α-nDCG, which is a function of "information nuggets" in the query and documents. They favoured novelty in the ranked result set, using the value of α, by penalising the documents if they contain the same information which is already present in the documents ranked higher. Radlinski et al. [67] used related queries to re-rank top-k Web search results for a query. They generate a set of related queries R, by using Web search engine query logs, for a query q. To get the top-k number of results

they take $\frac{k}{|R|+1}$ number of results from each query in R and from q. With this setup, they present a diverse set of results by assuming different user information needs to be present in related queries. Agrawal et al. [3] took into account the relative importance of different categories for the queries. They tried to produce diverse resulting documents based on relevance to each category, ranked by its relative importance, of the query.

The motivation for said existing studies was mainly to deal with the ambiguity of a user query or the multiplicity of a user intent — the central problem that their proposed techniques have attempted to solve is to find the optimal balance between the relevance and the diversity of search results [14, 93, 67, 3, 90, 79, 95, 96, 76]. The performance measures used in those papers, such as α-nDCG [21], combine both relevance and diversity. However, all these papers have only investigated the *topic diversity*, i.e. the coverage or variety of topics in the top-k search results. Besides topic diversity, there are other dimensions of diversity. For example Giunchiglia et al. [30] define these dimensions of diversity as: diversity of sources (multiplicity of sources of texts and images); diversity of resources (e.g. images, text); diversity of topic; diversity of viewpoint; diversity of genre (e.g. blogs, news, comments); diversity of language; geographical/spatial diversity; and temporal diversity.

The primary focus of our work in this thesis is to empirically analyse the diversity over the Web. The important questions to investigate the diversity over the Web are:

(i) Are Web search results dominated by major websites and therefore lacking diversity?

(ii) What is the topic diversity of Web search results?

(iii) Does advert websites on Web search engines also follow the same pattern as organic websites?

(iv) What is the diversity of queries coverage in Web search engines?

We aim to answer these questions by quantitatively modelling the diversity of:

(i) diversity of organic websites,

(ii) topic diversity in Web search results,

(iii) diversity of advert websites, and

(iv) diversity of queries coverage,

in two major search engines, Google and Bing, by making use of two diver-

sity measures well-studied in ecology, namely *Inverse Simpson's index* and *Shannon's diversity index*.

Thereafter we use the Cox proportional hazard regression model to investigate the factors which affect query popularity, i.e. the effect of diversity-related factors with which we can predict the number of months a query remains in the Google top charts.

We further investigate diversity in a domain other than Web search results, i.e. diversity of movies in genres or box office gross. The motivation behind this new dataset is that it has the additional time dimension and it also has overlapping categories since a single movie can be categorised into multiple movie genres.

After analysing diversity in search query results and movies, we apply these diversity measures to two application domains:

(i) to get the optimal number of balanced clusters for text data and

(ii) to produce a meta-evaluation for ranking of the top-k search query results that are relevant to the diverse needs of different users whose exact information needs are uncertain.

We note that all the programming in the experiments and analysis for our thesis has been performed in the Python language [85].

To understand and analyse the Web data in terms of diversity can help organisations better satisfy user needs, by providing information from different perspectives available in a wide range of sources in the Web. For example, if a Web search engine e.g. Google, can understand the relationship between diversity and user satisfaction, it can satisfy more users who visit their website, which will help grow their own business.

1.2 THESIS STRUCTURE

In Chapter 2 we describe the past and current literature about the work in diversity and its concepts as a whole and in particular over the Web. We discuss any connections with the work done concerning diversity over the Web and how our work is different from the work of others. In Chapter 3 we analyse and compare the diversity in Web search engines. We investigate the diversity in three different parts of Web search results page, i.e. diversity in organic Web search results, diversity in adverts presented alongside organic Web search

results and in the end, we analyse the topic diversity in Web search engines. In Chapter 4 we further investigate into the diversity of queries coverage in Web search engines. Afterwards, we analyse the role of diversity-related and other covariates affecting the popularity or lifetime of queries to remain in Google top charts. In Chapter 5 we take a look into additional dimensions of diversity. In particular, we analyse the diversity in movies related dataset which has overlapping categories. Afterwards, we also look into how the diversity is changing with respect to time. In Chapter 6 we investigate the role of diversity in clustering. We show how the diversity measures can be used to estimate the number of clusters in a given dataset. Afterwards, we show how to meta-evaluate the evaluation methods for ranking the diversified Web search results. Lastly, Chapter 7 summarises the thesis by describing the main contributions and prospects for the future research and development in the diversity over the Web.

CHAPTER 2

CRITICAL REVIEW

This chapter presents a critical review of the research and literature related to the diversity in the Web.

In Section 2.1 we shall talk about information retrieval in Web search engines and the importance of Web search engines in the Web. We also talk about how the search results in Web search engines are evaluated and presented to fulfil user requirements. After that in Section 2.2 we look into the diversity matrices that we propose to use in the Web to diversify the Web search results. Lastly in Section 2.3 we present the methods which are proposed in the literature, to diversify the Web search results.

2.1 WEB SEARCH AND INFORMATION RETRIEVAL

Many people use the Internet in everyday activities, and Web search engines have become an integral part of the Web to access information on websites. A survey conducted in 2012 has a finding that about 91% users use the Web search engines to find any required information over the Web. The primary focus of an IR (information retrieval) system in general and Web search engines, in particular, is to provide the results based on relevance to the user queries. On the other hand, the relevance of the retrieved documents is subjective based on the user who posted the query [11]. Web search engines provide the ranked results based on relevance to the query. The evaluation of relevance based ranking of Web search results has also been discussed in the research literature [53, 49, 26]. Vorhees [88] suggested a measure, mean reciprocal rank

(MRR), for evaluating the relevance of ranked results. It is basically the inverse of the rank of first relevant result for a query and then takes an average of this number for a certain number of queries in the system. It is defined as follows:

$$MRR = \frac{1}{Q} \times \sum_{i=1}^{Q} \frac{1}{r_i} \tag{2.1}$$

where Q is the number of queries and r_i is the rank or position of the first relevant result in the ranked Web search results.

K Järvelin and J Kekäläinen [42] proposed a method, normalised discounted commutative gain (nDCG), based on the graded relevance of documents presented in the ranked results. Graded relevance simply represents the relevance of a document to a particular query. For example, this relevance can be a number from 0 to 5, 0 for not relevant at all and 5 for perfectly relevant. nDCG is basically normalised DCG after dividing by ideal DCG. Whereas DCG is defined as:

$$DCG = \sum_{i=1}^{k} \frac{2^{GR_i} - 1}{\log_2(i+1)} \tag{2.2}$$

where k is the number of Web search results for a query and GR_i is the graded relevance of the document at position i.

While search engines are the primary source to find information over the Web, the users need to formulate a query in order to search for the required information. This has been observed through the research that the queries posted by users remain short and are limited to certain keywords hence these queries are not always clear and does not properly reflect the actual needs or requirements of the corresponding users [74, 81, 40, 41]. These type of queries which are not clear are termed as ambiguous queries. Queries posted by users on Web search engines are classified into three classes [21, 81]. Ambiguous, underspecified and clear queries. Ambiguous queries are the queries which have multiple interpretations or different meanings attached to it. For example, a query “jaguar” is an ambiguous query because it is not clear that required information by the user is about “jaguar animal” or “jaguar car” or is it about “jaguar novel”. Underspecified queries or semi-ambiguous queries, on the other hand, are those queries which have a clear interpretation but has multiple aspects or sub-topics, which are not yet clear. For example a

query “iPhone” has a clear interpretation, i.e. iPhone smartphone, but it is not clear what the user wants to know about iPhone, the user could either be interested in buying an iPhone or iPhone reviews or some other iPhone related news, e.g. an upcoming iPhone launch event. Finally, the clear queries are the queries which have a clear and well-understood meaning. For example “Birkbeck University of London”.

There has been a surge of research to deal with the ambiguous and underspecified queries [41, 81, 40, 75]. One possible solution is to diversify the Web search results based on multiple interpretations of a query, i.e. provide results from all possible interpretations or subtopics of the query. We discuss the general concept of diversity in Section 2.2. Afterwards, in Section 2.3, we review and explain the research literature concerning diversity in the context of Web and information retrieval.

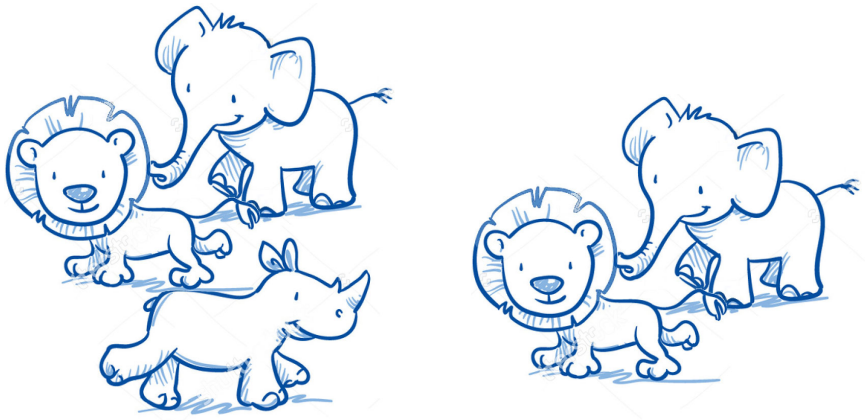
2.2 DIVERSITY MEASURES

The concept of diversity originated from ecology [52], has been widely diffused into many other scientific disciplines [82, 60] (such as linguistics and sociology).

When measuring the biological diversity of a habitat, where types of interest or concern are usually species, it is important to consider not only the number of different species present but also the relative abundance of each species. In the literature of ecology, the former is called *richness* and the latter is called *evenness* [62, 52, 10]. The measure of richness on its own cannot provide a full picture of diversity, as it does not account for the varying proportions of the number of individuals in any species. For example, intuitively, one wild-flower field with 500 daisies and 500 dandelions should be more diverse than another wild-flower field with 999 daisies and 1 dandelion — although they both have the same richness (two species), evidently the first field has much higher evenness than the second field.

The *richness*, R of the search result set for a query (topic) could be just defined as the number of distinct subtopics or interpretations appeared in the set.

For Example, in Fig. 2.1, it is shown that jungle a has three species, Lion, Elephant and Rhino, whereas, jungle b has only two species, Lion and Elephant. Hence jungle a has greater Richness ‘3’ than jungle b Richness ‘2’.



(a) Richness = 3 (b) Richness = 2

Figure 2.1: Differences in *Richness* in two jungles, a and b.

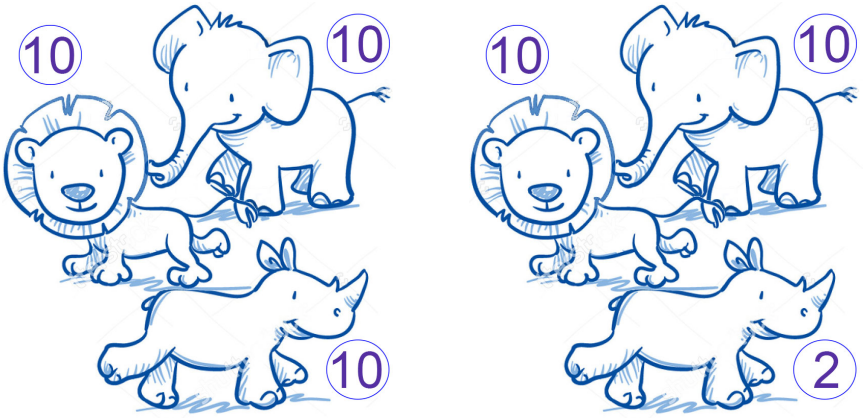
On the other hand, the *evenness* of the search result set for a query (topic) refers to how close in numbers each subtopic in the set is, i.e. it quantifies how evenly the search results are spread over the subtopics. For example, a search result set having 5 results from subtopic u and 5 results from subtopic v should have greater evenness than a search result set having 2 results from subtopic u and 8 results from subtopic v. In another example, from the species point of view, as shown in Fig. 2.2, jungle a has more evenly distributed species than jungle b. Hence jungle a has greater *Evenness*, '1' than jungle b *Evenness*, '0.8'.

Mathematically, the value of evenness is calculated as the normalised diversity:

$$evenness = D/D_{\max} \ , \tag{2.3}$$

where D is a diversity index, and $D_{\max}$ is the maximum possible value of D.

Although there exist many different diversity measures (such as HCDT entropy and Renyi entropy) and it is debatable which diversity index is the best [38, 43], we choose to use Inverse Simpson's Index [78], Shannon's Shannon's diversity index and Rao's quadratic entropy [69] to analyse the diversity in the Web, because the former two measures take into account both richness and evenness, and the latter one measure the sizes (richness) of species (groups) and the distances between species (groups).



(a) Evenness = 1 (b) Evenness = 0.8

Figure 2.2: Differences in *Evenness* in two jungles, *a* and *b*.

2.2.1 Inverse Simpson's Index

Inverse Simpson's index, aka Herfindahl-Hirschman index in economics [35, 37], is defined as

$$D = \left(\sum_{i=1}^{N} p_i^2 \right)^{-1}, \tag{2.4}$$

where N is the total number of different species (i.e., richness), and p_i is the proportional abundance of the i-th species (i.e., the proportion of the individuals belonging to the i-th species relative to the entire community of individuals in the area). It could be interpreted as the inverse of the probability that two individuals randomly selected belong to the same species. This index, which indicates the "effective" number of species [36], starts with 1 as its minimum possible value (representing a community containing only one species). For example when there are only lions in a jungle and no other species is available then Simpson's diversity index is 1 for that jungle. This index has its maximum possible value N (which occurs when all the N species are equally common in the community of interest). For example when there are lions and elephants each has 10 individuals then Simpson's diversity index is $N = 2$ for that jungle. The *evenness* under this index could be calculated as the normalised diversity index $D_E = D/N$ which ranges between 0 and 1.

2.2.2 Shannon's Diversity Index

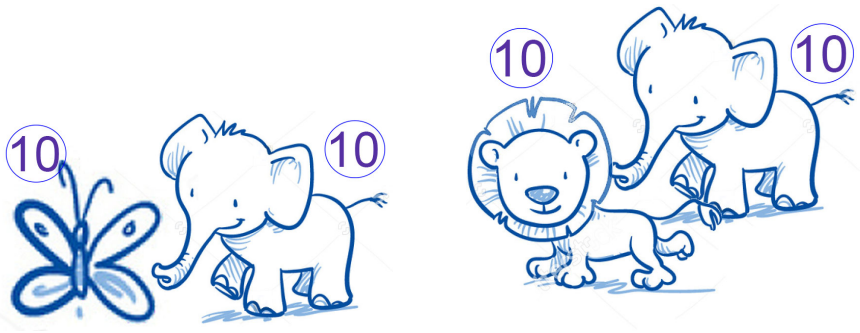
Shannon's diversity index [77], aka *entropy* in information theory [51], is defined as

$$H = -\sum_{i=1}^{N} p_i \ln p_i \ , \tag{2.5}$$

where N and p_i refer to the same quantities as in Section 2.2.1. It could be interpreted as the uncertainty (i.e., the degree of surprise) in predicting the species identity of an individual that is taken at random from the community of interest. This index has the minimum possible value 0 (which occurs when there is only one species). For example when there are only lions in a jungle and no other species is available then Shannon's diversity index is 0 for that jungle. The maximum possible value for this index is $\ln N$ (which occurs when all the N species are equally common in the community of interest). For example when there are lions and elephants each has 10 individuals then Shannon's diversity index is $\ln 2 = 0.69$ for that jungle. The *evenness* under this index could be calculated as the normalised diversity index $H_E = H/\ln N$, which ranges between 0 and 1.

Shannon's Diversity Index is known as Shannon's entropy in information theory. It is the core concept in information theory. It was proposed by Claude E. Shannon [77], who is known as the father of information theory. Shannon's entropy, in general, refers to disorder or uncertainty in predicting the outcome of an event. For example, when we toss a fair coin then each side, heads and tails, has an equal chance or probability to appear as a result. Hence $H = 1$ when p_i, in this case, is 0.5 for both the outcomes. Which is the maximum of Shannon's entropy for this system (coin) i.e. it is maximum uncertainty in predicting the outcome of a coin. Now consider a coin which is not a fair coin e.g. heads has a greater chance, 0.8, to appear in the outcome than tails, 0.2. For this coin which is not fair, Shannon's entropy is 0.72. It is because in this unfair coin there is less uncertainty in predicting its outcome. For the coin which is the same on both sides e.g. both sides tails, shall have a zero Shannon's entropy because there is zero uncertainty in predicting its outcome.

These two diversity measures discussed in Section 2.2.1 and Section 2.2.2, both being popular in the literature of ecology and other scientific disciplines such as linguistics [6] and sociology [4], can be regarded as two special cases of the general entropy function $\left(\sum_{i=1}^{N} p_i^{\alpha}\right)^{\frac{1}{1-\alpha}}$: when $\alpha = 2$ it is Simpson's



(a) Higher Differences (δ) (b) Lower Differences (δ)

Figure 2.3: Differences in between species in two jungles, a and b.

index and when α approaches 1 it becomes Shannon's diversity index [60]. It is debatable which diversity index is the best [38, 43]. In this paper, we make use of both.

2.2.3 Rao's Quadratic Entropy

Along with *richness* and *evenness*, it is also important to take into account the measure of distance or difference between two species, to measure diversity. For example, as shown in Fig. 2.3, intuitively, an animals field, jungle a, with 10 elephants and 10 butterflies should be more diverse than another animals field, jungle b, with 10 elephants and 10 lions — although they both have the same richness and same evenness, evidently the first field has a much higher difference or distance than the second field.

For this purpose, we use *Rao's quadratic entropy* [69] as it takes into account both the sizes of species and the distances or differences between species. Rao's quadratic entropy, denoted by Q, is given by

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j \delta(i,j) \ , \tag{2.6}$$

where N is the number of species and p_i, p_j is the proportions of species i and j, respectively, and $\delta(i,j)$ is the difference or distance between them.

2.3 DIVERSITY IN INFORMATION RETRIEVAL

In recent years, a variety of quantitative measures of diversity have been successfully applied in computer science for Web search [21, 28, 45, 74, 96], text

mining [7], and recommender systems [15].

If the queries that are posted by users are either ambiguous or underspecified, apparently the search results, which are not diversified, provided by a Web search engine, shall not satisfy the requirements of all of its users. On the other hand, if the search results are diversified over multiple interpretations or sub-topics of the users' queries then there is greater chance that the user shall be satisfied with at least one result from the search results.

Radlinski et al. [66] classify the diversity approaches in two separate classes as either extrinsic or intrinsic. Extrinsic approach deals with users' information needs when the queries are ambiguous. Whereas intrinsic approaches explore the ways to deal with redundancy in the result set when the queries are underspecified or semi-ambiguous.

A number of approaches have been discussed in the literature to deal with diversifying the Web search results.

Carbonell and Goldstein [14] were among the earlier authors to introduce the diversity in a result set. They introduced Maximal Marginal Relevance (MMR) which not only use relevance between documents and a query but also take into account for similarities between document to document to reduce the redundancy and introduce novelty. They get the ranked list of document set by maximising:

$$MMR = \arg\max_{d \in D} \left[\alpha(sim_{QD}(d, q) - (1 - \alpha) \max_{d_j \in S} sim_{DD}(d, d_j)) \right] \tag{2.7}$$

where D is the set of documents returned by a Web search engine, $sim_{QD}(d, q)$ is the similarity between query q and document d and represents the relevance, $sim_{DD}(d, d_j)$ is the similarity between document d with other documents in the result set S which are ranked higher than document d and parameter α controls the balance between relevance and novelty, a measure of diversity, of the result set.

Chapelle et al. [18] proposed an extension to the Reciprocal Rank for the case of graded relevance and described Expected Reciprocal Rank, ERR, for graded relevance to evaluate the ranking of the Web search results. This method is based on the assumption that usefulness of the document at rank i is not independent from the usefulness of documents before rank i, which is not the case in nDCG [42] in which document's relevance or usefulness is

independent of each other. While in DCG works only with graded relevance, which is independent of other documents in the result set, ERR, on the other hand, takes into account for the probability of user satisfaction at every position of ranked list and discounts the measure accordingly. ERR is defined as:

$$ERR = \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r$$

where n is the total number of documents in ranking, r is the rank of the document and $\prod_{i=1}^{r-1} (1 - R_i) R_r$ is the probability that user stops at position r, and $(1 - R_i)$ is the discount on R_r.

$$R_i = \mathcal{R}(g_i)$$

where $\mathcal{R}(g_i)$ is defined as mappings from relevance grades to probability of relevance which is chosen in accordance to gain function of DCG as:

$$\mathcal{R}(g_i) = \frac{2^g - 1}{2^{g_{max}}}, \quad g \in \{0, ..., g_{max}\}$$

where g is the graded relevance of the document.

Clarke et al. [21] studied diversity and novelty in the answers to any question. They focused on ranking the documents containing answers, "information nuggets", to a question with respect to relevance and diversity which reduces redundancy in favour of novelty. They proposed a generalised version of the function for nDCG [42] and named it α–nDCG which is a function of information nuggets in question and its answers.

For their experiments, they used the dataset from the TREC 2005 question answering task [89]. In this, they were given a topic, query, and a set of questions related to the topic. By using the topic as query their goal is to provide answers to these questions from the corpus of the collection of newspaper articles used at TREC.

Basically, they considered a document to be ranked highest when it answers the highest number of questions. If two documents answer the same number of questions then the document which is not answered before is ranked higher. In this way, they introduced novelty in the result set.

They defined the probability of a document d ranked at position k as:

$$P(R_k = 1|u, d1, ...) = \gamma\alpha \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_i,k-1} \tag{2.8}$$

where u is the information need occasioning a user to formulate query, i is the information nugget, d_k is the document at rank k, γ is the probability of i being in u, $P(i \in u)$, which is assumed to be constant for all i, $J(d_k, i)$ is the human judgement for whether nugget i is present or not in document k and it is either 0 or 1, α is a constant, with $0 < \alpha \leq 1$, which controls the novelty in results by penalizing a document d_k if it contains the same information nugget as in d_{k-1} documents, $(r_i, k-1)$ is the number of documents ranked up to position $k-1$ that have been judged to contain nugget ni which is given as:

$$r_i, k-1 = \sum_{j=1}^{k-1} J(d_j, i)$$

By dropping the constant $\gamma\alpha$ from Eq. (2.8), which has no impact on relative values, they define the kth element of the gain vector G, for α–nDCG, as:

$$G[K] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_i,k-1} \tag{2.9}$$

By using the gain vector defined by Eq. (2.9) they calculated nDCG measure. Due to the importance of α in it, they call it α–nDCG. When the value of α is 0 then α–nDCG correspond to standard nDCG measure.

Bache et al.[7] used an approach which uses the content of a document to quantify its diversity. Their focus is to compute the diversity of each document relative to the rest of the corpus. For this task, they used Rao's quadratic diversity [69] which is defined in Eq. (2.6).

They create a $D \times T$ document-topic count matrix, using LDA topic model, with entries n_{dj} corresponding to the number of word tokens in document d that are assigned to topic j. Using this they define the diversity per document as:

$$div^{(d)} = \sum_{i=1}^{T} \sum_{i=1}^{T} p(i|d)p(j|d)\delta(i, j) \tag{2.10}$$

where $P(i|d)$ is the proportion of word tokens in document d that are assigned to topic i, and $\delta(i, j)$ is a measure of the distance between topic i and topic

j. For the experiments, they used three different types of datasets. First is the PubMed Central Open Access dataset which is comprised of articles published in biomedical journals. Second is the NSF Awards from 2007 to 2012 gathered from www.nsf.gov/awardsearch and the third dataset used is the Association of Computational Linguistics Anthology Network consisting of papers published in selected computational linguistics conferences.

Agrawal et al. [3] took into account the relative importance of different categories for the queries and documents in Web search. They used the probability distribution for categories C of a query q, $P(c|q)$, i.e. probability of a given query belonging to different categories from ODP taxonomy. Along with that they also used the quality value of a document d for query q when the intended category is c, $V(d|q,c)$. This can be interpreted as the probability that a document d satisfies the information need having query q in category c. To find a set of documents S from the corpus D they used:

$$P(S|q) = \sum_{c} P(c|q)(1 - \prod_{d \in S}(1 - V(d|q,c))) \tag{2.11}$$

where (1-V(d—q,c)) is the probability that the document d does not qualify to fulfil the information need for given query q in category c. Hence the Eq. (2.11) will provide at least one document to satisfy user's information intentions for the query q using its relative importance in different categories.

Radlinski et al. [68] used user clicks to learn the best ranking for results in order to present a diverse set of documents. For a fixed query q, they go through all the documents by considering every position, for top-k positions, for every document and learn the ranking of documents from different users' clicks. By considering user clicks from all the users who have a diverse set of requirements for the same query, the ranking algorithm implicitly takes care for the result set to become diverse.

In order to diversify top-k Web search results for query q, Radlinski et al. [67] used related queries to rerank them. They generate a set of related queries R, using query logs from a Web search engine, for query q. To get the top-k number of results they take $\frac{k}{|R|+1}$ number of results from each query in R and from q.

To evaluate the diversity they use relevance feedback $match(d_i, u)$, how well document d_i matches the interest of user u. In order to get the diversity

score across a set of users U they take an average of the maximum $match$ score, in document set D, for all the users as $\frac{1}{|U|}\sum_{(u \in U)} \max_{d_i \in D} match(d_i, u)$.

CHAPTER 3

WEBSITES DIVERSITY IN WEB SEARCH ENGINES

Web search engines nowadays are the primary means for people to locate information over the Internet [49]. A survey conducted several years ago reported that about 91% of users would employ search engines for their information needs [65]. It has also been observed that users' clicks on search results are heavily biased towards those on the top of the results list [25]. Many people believe that if a website is not in the top search results of a mainstream search engine, it has minimal influence and effectively does not exist. Hence it is important to investigate the website (source) diversity of Web search engines.

In this chapter, we set out to investigate to what degree Web search results are dominated by major websites (such as Amazon), i.e. how diverse Web search results are. In addition to the *organic* search results that are returned by a search engine to a user because of their relevance to the corresponding search query, we also look into *sponsored* search results that appear as advertisements.

For a similar reason, that diversity is crucial to the sustainability of ecosystems and the prosperity of any human society [60], a healthy level of diversity is very important for the advancement of the Web. Lack of diversity in Web search engines may imply that small, new websites do not have a fair chance to compete with large, old well-established websites, and thus users are limited to a narrow choice of information channels.

The rest of this chapter is organised as follows. Firstly, we analyse the diversity of organic websites in two Web search engines, with two measures

of diversity i.e. Inverse Simpson's diversity index and Shannon's diversity index along with their corresponding *richness* and *evenness*. Secondly, we have an insight into the topic diversity of Web search results and see how the different interpretations i.e. sub-topics, are covered in the main query results. Thereafter, we examine the diversity in adverts websites that appear along with the main Web search results. Afterwards, we perform a randomised significance test to see if the changes in the diversity values in two Web search engines are significant or not. Lastly, we draw conclusions for this chapter.

3.1 DIVERSITY OF ORGANIC WEBSITES

3.1.1 Overview

Are Web search results usually dominated by major websites and therefore lacking diversity? We aim to answer this question by quantitatively modelling the diversity of search results for popular queries by using two diversity measures well-studied in ecology, namely *Inverse Simpson's diversity index* and *Shannon's diversity index* (see Section 2.2).

3.1.2 Presentation and Analysis of Data

To collect typical Web search results data for our investigation, we first gathered all the popular queries over 114 months from January 2004 until June 2013 in the six representative categories of Google Top Charts, and then downloaded the top-k ($k = 10$ and $k = 50$) organic search results as well as all the sponsored search results for those queries from two mainstream Web search engines with most users: Google and Bing. The six categories are: (i) Shopping, (ii) Nature & science, (iii) Sports, (iv) Business & politics, (v) Travel & leisure, (vi) Entertainment.

Fig. 3.1 shows Top Charts queries for January 2004, in two categories, namely, Travel & leisure and Nature & science. All the queries are pre-categorized by Google, into six different categories. It can also be seen that all these queries fall into the definition of underspecified queries [74] i.e. these queries are neither completely ambiguous nor clear.

Since we focus on measuring website diversity, we extract the hostname from each search result's corresponding URL as its website address. For ex-

```
<TopChartQueries>
  <Date value="Jan 2004">
    <Category MainCat="Travel & leisure">
      <Query>Wine</Query>
      <Query>Coffee</Query>
      <Query>Cake</Query>
      <Query>Pizza</Query>
      <Query>Chocolate</Query>
      <Query>Cookie</Query>
      <Query>Soup</Query>
      <Query>Beer</Query>
      <Query>Chicken</Query>
      <Query>Tea</Query>
    </Category>
    <Category MainCat="Nature & science">
      <Query>Dog</Query>
      <Query>Cat</Query>
      <Query>Horse</Query>
      <Query>Fish</Query>
      <Query>Bird</Query>
      <Query>Bear</Query>
      <Query>Chicken</Query>
      <Query>Cow</Query>
      <Query>Monkey</Query>
      <Query>Rabbit</Query>
    </Category>
  </Date>
</TopChartQueries>
```

Figure 3.1: Top chart queries, for Jan 2004, in two categories, i.e. “Travel & leisure” and “Nature & science”.

Table 3.1: The dataset of organic search results.

			Google	Bing
top-k	category	#queries	#results	#results
10	Shopping	106	990	985
	Nature & science	346	3570	3539
	Sports	746	7432	7421
	Business & politics	397	4350	4273
	Travel & leisure	539	5660	5634
	Entertainment	1625	16215	16037
50	Shopping	106	4866	2998
	Nature & science	346	17233	10635
	Sports	746	35919	26122
	Business & politics	397	19766	14325
	Travel & leisure	539	26911	17306
	Entertainment	1625	77814	51696

Table 3.2: The websites that appear most often in the top-50 organic search results.

	Google		Bing	
rank	website	freq.	website	freq.
(1)	Wikipedia	4454	Wikipedia	5735
(2)	Youtube	3775	Imdb	3343
(3)	Amazon	3629	Youtube	3037
(4)	Facebook	3271	Amazon	2501
(5)	Google Images	3221	Twitter	1395
(6)	Twitter	2575	Facebook	1294
(7)	Imdb	2475	HuffingtonPost	1251
(8)	Google Sites	1905	Bing Images	1210
(9)	TheGuardian	1657	Bing Videos	1134
(10)	DailyMail	1409	Espn	1020

ample, the website address of the URL `www.acm.org/sigs/publications/` would be just "`acm.org`".

Table 3.1 shows the number of search queries and the number of organic search results pooled together for each category in this real-world data set.

One phenomenon that we can immediately observe from the data is the

high skewness of the websites' distribution — a few popular websites (see Table 3.2) occur very frequently, while the majority of websites occur only a small number of times. The websites that appear most often in the top-50 organic search results (across all the six categories) from Google and Bing are listed in Table 3.2. The table of the most popular websites in the top-10 search results look quite similar, so it is omitted.

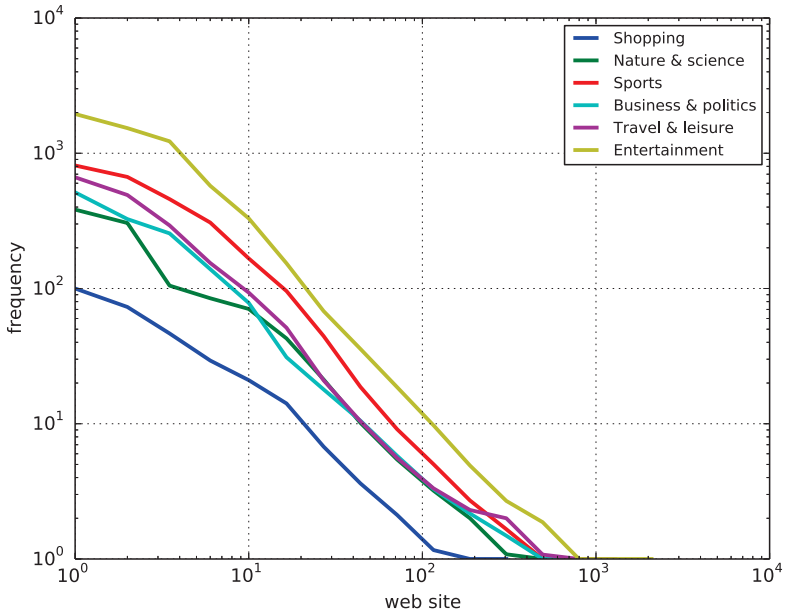
Figs. 3.2 and 3.3 shows the log-log plots of website frequency (against rank order) in the top-10 and top-50 organic search results respectively, where the curves have been smoothed using the *Fibonacci binning* [86]. Fibonacci binning is a simple logarithmic binning technique in which bins are sized like the Fibonacci numbers. It makes it visually more accurate than power-of-b binning (where $b = 2, 10$) [87].

As we have anticipated, all of those log-log plots are roughly in the shape of straight lines, which indicates that the distribution of websites follows *Zipf's law* [64] — the frequency of the i-th popular website, f_i, is proportional to $1/i^s$, where s is the exponent characterising the distribution (shown as the slope of the straight line in the corresponding log-log plot). It is known that Zipf's law holds if the number of occurrences of each element is independent and identically distributed random variables with power law distribution [2]. As power law is prevalent on the Internet [58], it is not surprising to see that the distribution of website in the top-k search results can be well modelled by Zipf's law. This could probably be explained by an underlying preferential attachment process (i.e. "the rich get richer" phenomenon) [9].

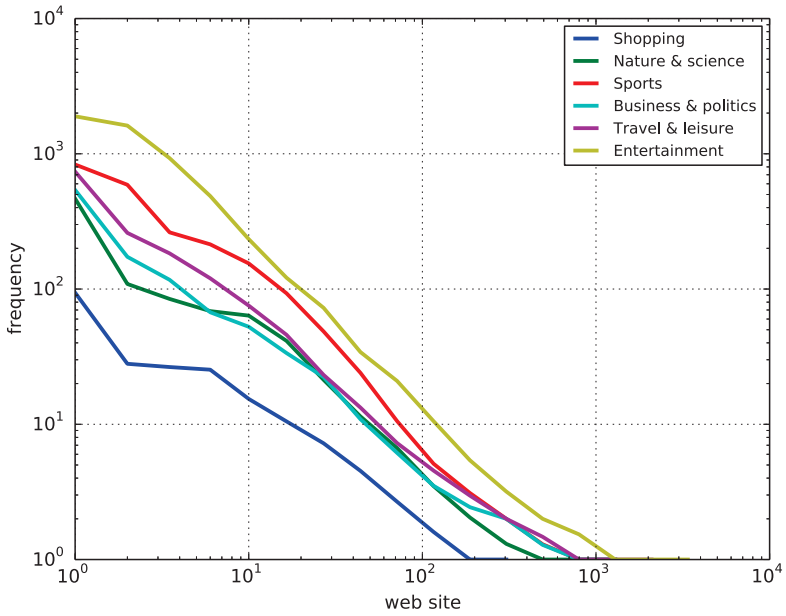
Without loss of generality, we could rank all the N distinct websites in the top-k organic search results according to their frequencies, then the proportion of the i-th website would be given by

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_j} = \frac{1/i^s}{\sum_{j=1}^{N}(1/j^s)} \,. \tag{3.1}$$

Table 3.3 shows the number of distinct websites N together with the Zipfian exponent s for the top-k organic search results in each category from Google and Bing, where s was estimated by employing linear regression to fit the log-log plot. The coefficient of determination for the regression, R^2, is greater than 0.96 for all the categories, which means that the data fit the model of Zipf's law very well. In our data set, s is between 0.7 and 1.2.



(a) Google

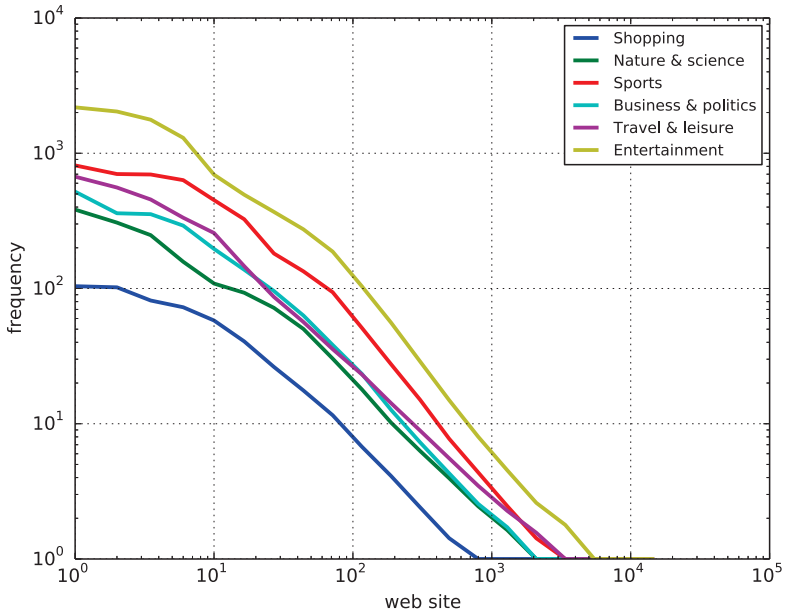


(b) Bing

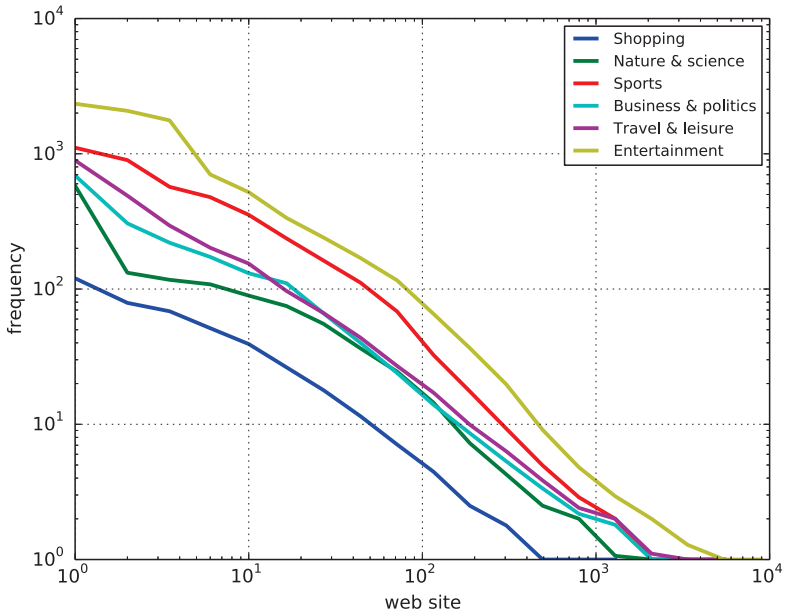
Figure 3.2: The log-log plots of website frequency in the top-10 organic search results.

3.1.3 Theoretical Analysis of Diversity

As the occurrence of each distinct website in the search results is, in general, governed by Zipf's Law Eq. (3.1), we should be able to analytically compute the two diversity measures, Inverse Simpson's index D and Shannon's diversity index H using Eq. (2.4) and Eq. (2.5), respectively. For example, in the case of organic search results which follow Zipf's law, it can be shown that



(a) Google



(b) Bing

Figure 3.3: The log-log plots of website frequency in the top-50 organic search results.

Inverse Simpson's index $D = G^2_{N,s}/G_{N,2s}$, where $G_{n,m} = \sum_{i=1}^{n}(1/i^m)$ is the generalised harmonic number [46]. The corresponding evenness could also be obtained accordingly.

Hereafter we elaborate on the relationship between the aforesaid diversity measures and the Zipfian distribution for organic search results. It can be extended straightforwardly to the Zipfian distribution with the exponential

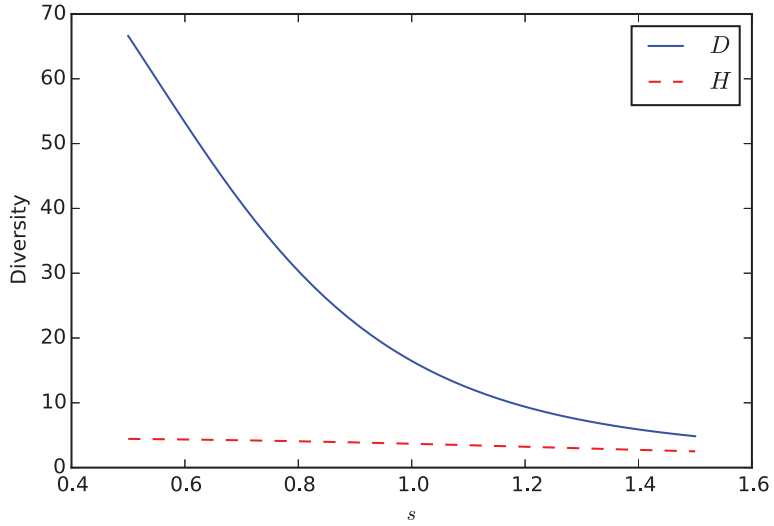
Table 3.3: The distribution of websites in the organic search results.

		Google		Bing	
top-k	category	#website N	exponent s	#website N	exponent s
10	Shopping	261	0.919	348	0.788
	Nature & science	915	1.004	1010	0.896
	Sports	853	1.166	1282	1.066
	Business & politics	1080	1.004	1475	0.905
	Travel & leisure	1582	1.042	1802	0.924
	Entertainment	2223	1.167	3120	1.077
50	Shopping	1751	0.739	1039	0.761
	Nature & science	7217	0.772	3561	0.794
	Sports	6689	0.910	4565	0.944
	Business & politics	6469	0.849	4916	0.825
	Travel & leisure	11020	0.809	6222	0.850
	Entertainment	16467	0.950	10476	0.974

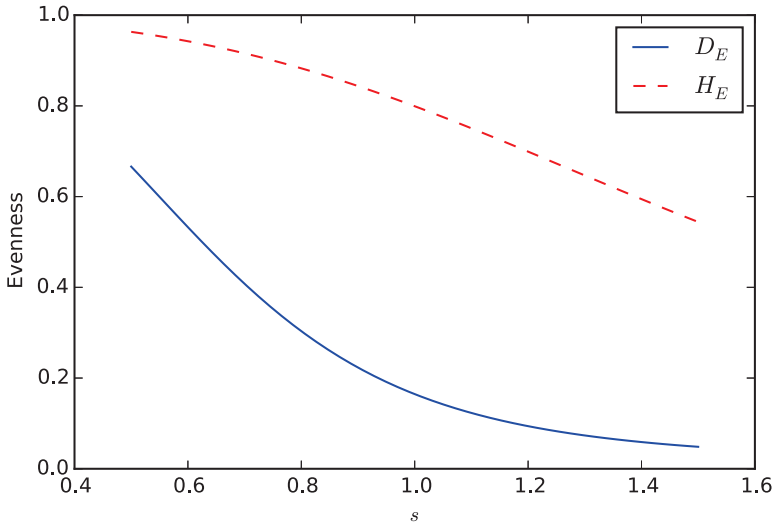
cut-off for sponsored search results see Section 3.3.1.

Fig. 3.4 shows how the website diversity and evenness are affected by the Zipfian exponent s when the number of distinct websites N, i.e. the website richness, is fixed. It can be seen from Fig. 3.4 that as s increases, the website diversity, D or H, keeps decreasing, though the speed of decrease becomes slower, (i.e. the marginal loss is diminishing). Since richness does not change here, the sole reason for the decrease of diversity is just the decrease of evenness: when s is bigger, the Zipfian distribution of websites is more skewed, and thus there is less evenness as represented by D_E or H_E and consequently less diversity.

Fig. 3.5 shows how website diversity and evenness are affected by the number of distinct websites N, when the Zipfian exponent s is fixed ($s = 1.00$ in our case). It can be seen that as N increases, the website diversity, D or H, keeps increasing, though the rate of increase becomes slower, (i.e. the marginal return is diminishing). What is most interesting here is the relationship between the website richness N and the website evenness (as shown in Fig. 3.5b). For a given Zipfian distribution of websites with a specific Zipfian exponent s, it turns out that the website evenness, D_E or H_E, actually decreases as the website richness N increases. This implies that there is a tension



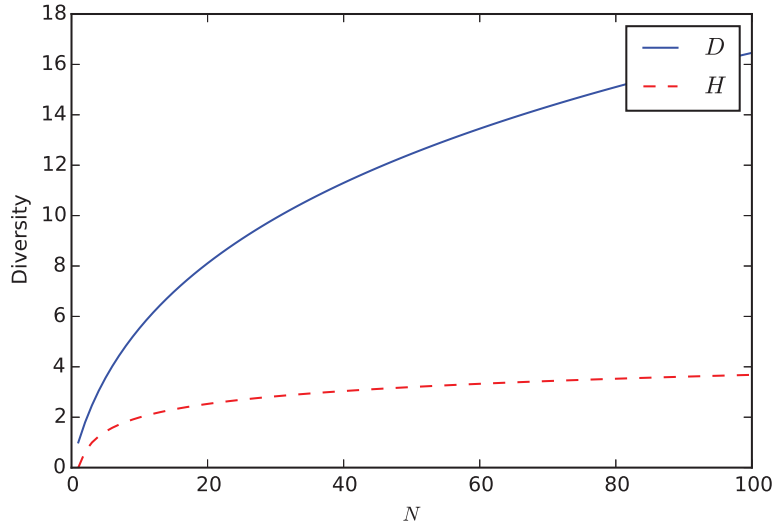
(a) diversity



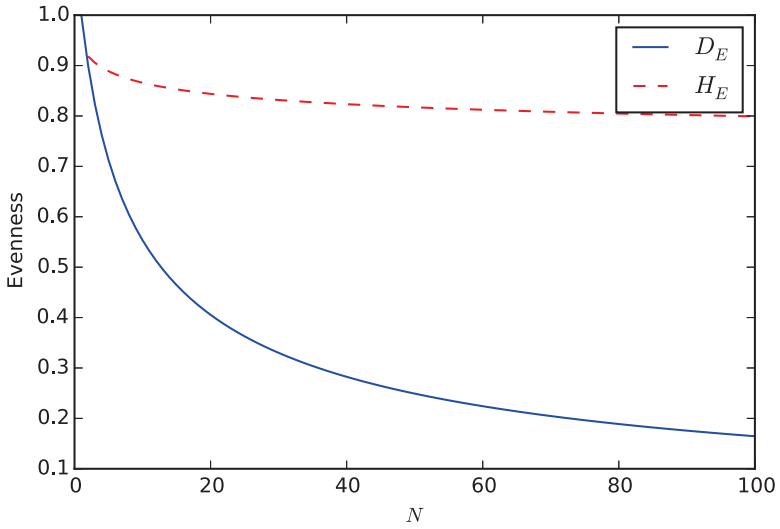
(b) evenness

Figure 3.4: How the website *diversity* and *evenness* change with s (when $N = 100$).

between the website richness N and the website evenness E, assuming that the search results in one category have an intrinsic Zipfian exponent s (see Table 3.3). Nevertheless, the website richness seems to play a more important role than the website evenness here, because the overall diversity would still be higher with a bigger N (as shown in Fig. 3.5a), even though it reduces the website evenness. Zipf's law is known to generate a "long tail" distribution where a small high-frequency population is followed by a large low-frequency population which gradually tails off: although the low-frequency items at the long tail each has a low probability of occurrence, the total number of their



(a) diversity



(b) evenness

Figure 3.5: How the website *diversity* and *evenness* change with N (when $s = 1.00$).

occurrences could be bigger than that of the high-frequency items [5]. The above analysis suggests that when the long tail of websites becomes longer, the overall diversity of Web search results will be better due to the higher richness implied by a larger N even if the evenness might be lower. It has been found by previous studies that in online book sales [12] and consumer software downloads [94] the long tail has grown longer over time.

Table 3.4: The websites evenness of the organic search results.

		Google		Bing	
top-k	category	Simpson's D_E	Shannon's H_E	Simpson's D_E	Shannon's H_E
10	Shopping	0.136	0.806	0.166	0.863
	Nature & science	0.042	0.750	0.041	0.778
	Sports	0.031	0.655	0.027	0.698
	Business & politics	0.029	0.719	0.031	0.799
	Travel & leisure	0.020	0.716	0.022	0.769
	Entertainment	0.011	0.608	0.009	0.652
50	Shopping	0.125	0.866	0.137	0.865
	Nature & science	0.053	0.862	0.052	0.843
	Sports	0.029	0.767	0.028	0.753
	Business & politics	0.038	0.821	0.038	0.837
	Travel & leisure	0.028	0.847	0.028	0.835
	Entertainment	0.012	0.746	0.012	0.739

Table 3.5: The website diversity of the organic search results.

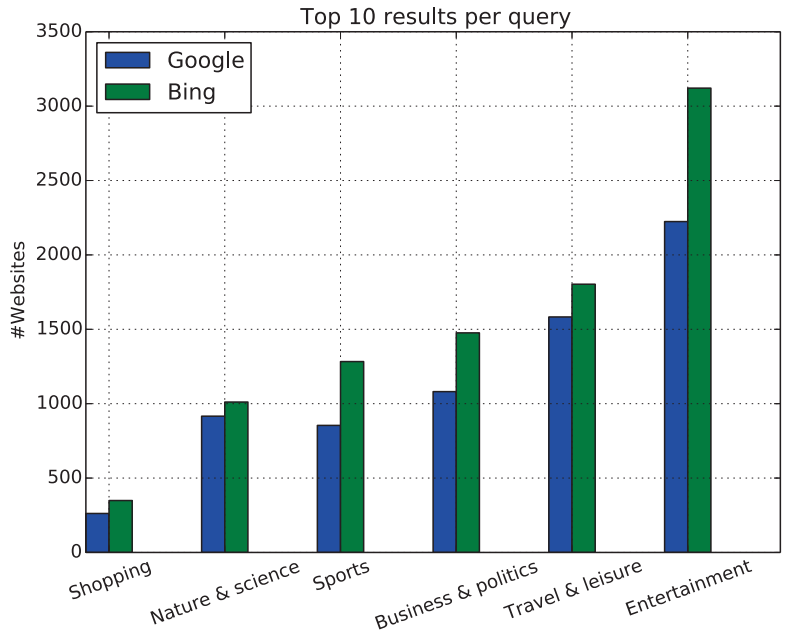
		Google		Bing	
top-k	category	Simpson's D	Shannon's H	Simpson's D	Shannon's H
10	Shopping	35.411	1.948	57.810	2.192
	Nature & science	38.015	2.222	41.149	2.338
	Sports	26.541	1.921	35.141	2.171
	Business & politics	30.826	2.182	45.702	2.531
	Travel & leisure	31.680	2.289	40.194	2.502
	Entertainment	23.728	2.035	27.768	2.277
50	Shopping	219.431	2.810	142.825	2.610
	Nature & science	384.124	3.327	183.509	2.993
	Sports	194.646	2.933	127.363	2.755
	Business & politics	246.298	3.128	188.837	3.088
	Travel & leisure	311.872	3.422	175.483	3.167
	Entertainment	198.295	3.144	120.601	2.972

3.1.4 Experiments and Results for Diversity in Organic Websites

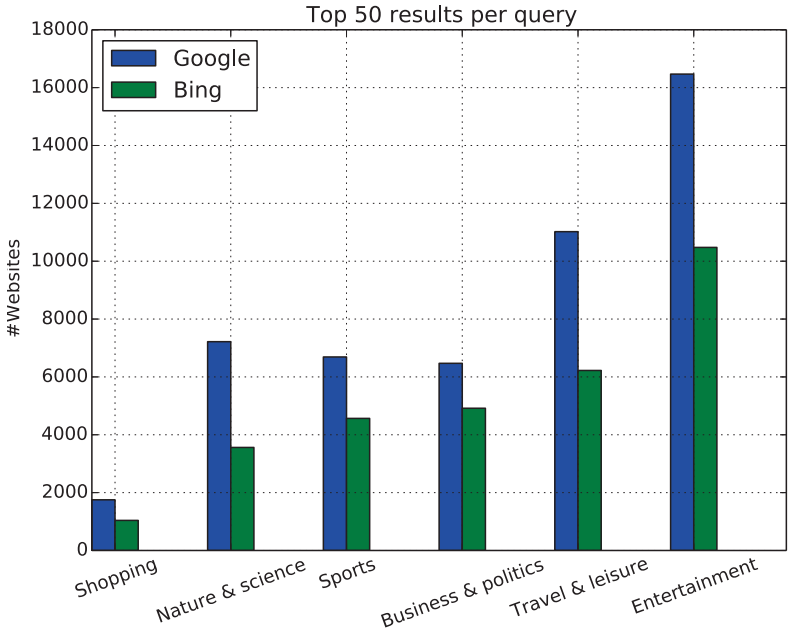
Currently, Google and Bing are the two major Web search engines for English users [1].

There arises the question: *How do they compare against each other in terms of the diversity of their search results?*

While analysing and examining the diversity of Web search engines it is also



(a) top-10



(b) top-50

Figure 3.6: The websites richness of the organic search results.

important to analyse the richness and evenness along side the Simpson's and Shannon's diversity index. Richness and evenness will help better understand, for example, in the diversity results for two Web search engines when the first is better in terms of richness but poorer in terms of evenness than the second Web search engine. Hence to analyse and better understand the differences we compare the search results in Web search engines in terms of Simpson's

and Shannon's diversity indices and their corresponding richness, evenness measures.

Fig. 3.6 shows the website richness of the top-k organic search results from Google and Bing (see also Table 3.3). An apparent pattern across all categories in this bar chart is that in the top-10 organic search results Bing's richness is consistently higher than Google's, but in the top-50 organic search results, Google's richness is consistently higher than Bing's. Such a discrepancy between these two search engines is probably rooted in the differences between their proprietary webpage ranking algorithms [47].

Table 3.4 shows the website evenness of the top-k organic search results for Google and Bing. Unlike the website richness, it turns out that Google and Bing do not exhibit a marked difference with regards to the website evenness: their evenness scores are very close to each other for any particular category: the largest gap between their Simpson's evenness D_E is 0.03; the largest gap between their Shannon's H_E is 0.06 (Shopping category). Moreover, there is no clear winner on the website evenness: Google has a higher evenness for some categories, but Bing has a higher evenness for some other categories. The website evenness for organic search results seems to be mostly determined by the category rather than by the search engine: it varies greatly from category to category, but does not change much when switching from one search engine to the other.

Since Google and Bing are roughly equivalent with regards to the website evenness, their overall diversity levels for organic search results would depend on the website richness. It can be seen clearly from Table 3.5 that in the top-10 organic search results Bing's diversity is consistently higher than Google's, but in the top-50 organic search results, Google's diversity is consistently higher than Bing's — the same pattern that the website richness follows. Here we can say, as it is evident from diversity and in particular richness, that compared to Bing, the Google presents less sources in top-10 Web search results results but it focuses on showing more sources (websites) if users go beyond the first page (top-50) of Web search results.

Since sites like Wikipedia, Imdb and DailyMail are highly reliable and authoritative in their relevant categories. Therefore people are more likely to prefer seeing results from them. This requires a further study into what level

of search results diversity is good to improve the user experience. Therefore, having a more diverse set of results should not be considered better or worse without considering the relevance and user preferences for any authoritative websites.

3.2 TOPIC DIVERSITY

Here we investigate the topic diversity in Web search results. When a user submits an ambiguous query, the Web search engines should present the search results from all the possible aspects or subtopics of the given query. There are several ways to get different possible aspects that are related to an ambiguous query. For example, a query expansion technique [70] can be used for this purpose which utilises the search results retrieved for the main query to redefine an ambiguous query. Web search engines could analyse the query logs to formulate the subtopics for an ambiguous query [8, 27]. Santos et al. [75] showed how the "related queries", shown along with the search results in Web search engine's interface, can be used to rank the diversifies search results. As the two Web search engines, Google and Bing, under investigation here, use the related queries to deal with queries having multiple aspects, we use these related queries as subtopics for the main query. After having the queries and their related queries, subtopics, we analyse the topic diversity in two Web search engines, Google and Bing.

3.2.1 Presentation and Analysis of Data

To analyse the topic diversity per query, we use the Top Chart queries, in six different categories: (i) Shopping, (ii) Nature & science, (iii) Sports, (iv) Business & politics, (v) Travel & leisure, (vi) Entertainment.

We also collect the related queries for these top chart queries from Web search engines. We collect the related queries, for any top chart query, from the respective Web search engine. For example related queries, for the main query "Barack Obama", are collected separately from Bing and Google Web search engines. Afterwards, we collect top-K search results for top chart queries and related queries from Google and Bing. Table 3.6 shows the number of queries and related queries, along with the average number of queries per query, across all the categories in Google and Bing.

Thereafter, we collect host-names of the URLs from these Web search result's snippets as we showed in Section 3.1.2. For example, the website address of the URL `http://www.acm.org/sigs/publications/` would be just "`acm.org`".

As shown in Section 3.1.2 that top chart queries are underspecified queries [74], i.e. these queries are neither completely ambiguous nor clear. Fig. 3.7 shows a query and its related queries. It can be seen that with the help of related queries (subtopics), for the top chart queries (topics), Web search engines are trying to clearly explain the intent of a user or different aspects of a query. Hence we assume that Web search engines use the related queries to reduce ambiguity in the main query.

```
<Query topic ="barack obama">
   <RelatedQuery>barack obama biography</RelatedQuery>
   <RelatedQuery>barack obama education</RelatedQuery>
   <RelatedQuery>barack obama parents</RelatedQuery>
   <RelatedQuery>barack obama sr</RelatedQuery>
   <RelatedQuery>barack obama Twitter</RelatedQuery>
   <RelatedQuery>barack obama facts</RelatedQuery>
   <RelatedQuery>barack obama net worth</RelatedQuery>
   <RelatedQuery>barack obama daughter</RelatedQuery>
</Query>
```

Figure 3.7: A query and it's related query

For the purpose of our analysis, topic diversity in Web search results, the related queries serve as the subtopics for a query topic. To consider a related query as a distinct subtopic we assume that all the related queries, for any main query, provided by Web search engines are distinct from each other in terms of information required from Web search engines.

To make an analogy to diversity of species in the field of Ecology, an intersection of number of unique URLs in related queries and a query, i.e. the URL that is present in the result set of a related query is also in the query, will be analogous to the total number of organisms or individuals of a particular species. For example if five URLs from a related query, subtopic, "Brack Obama Biography" are also present in the main query "Barack Obama" then the frequency (the number of individuals of any species) of this subtopic is five.

Table 3.6: Dataset for topic diversity, showing number of queries and their related queries (average number of related queries per query) across all categories

		Google	Bing
Category	#Queries	#Related Queries	#Related Queries
Shopping	106	810 (7.6)	809 (7.6)
Nature & science	346	2725 (7.9)	2718 (7.9)
Sports	746	5769 (7.7)	5745 (7.7)
Business & politics	397	3133 (7.9)	3124 (7.9)
Travel & leisure	539	4124 (7.6)	4116 (7.6)
Entertainment	1625	12446 (7.7)	12414 (7.6)

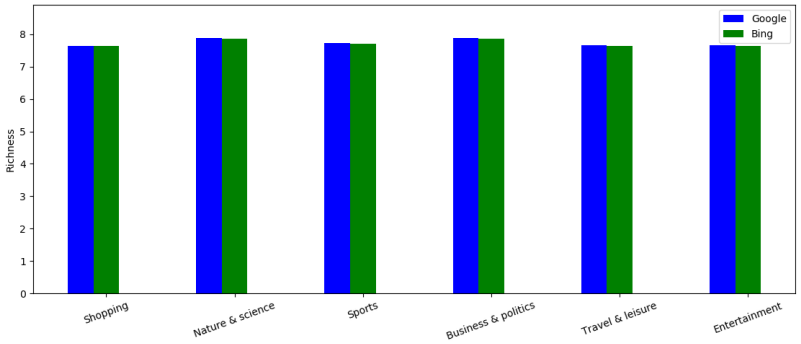


Figure 3.8: Average *richness* for every query in search results.

3.2.2 Experiments and Results for Topic Diversity

Fig. 3.8 shows the average topic richness, i.e. the average number of related queries per query, across all the categories, in Google and Bing (see Table 3.6). It shows that Google and Bing has somewhat the same richness, i.e. around eight number of subtopics per topic. This can be due to their subtopic formulation module used by these Web search engines.

On the other hand Table 3.7 shows the average *topic evenness* in Web search results from Google and Bing. It is shown here that Google has consistently higher evenness than Bing, in top 10 and top 50 search results, across all six categories. Which shows that Google has more balanced, evenly dis-

Table 3.7: Average topic evenness per query in Web search results.

		Google		Bing	
top-k	category	Simpson's D_E	Shannon's H_E	Simpson's D_E	Shannon's H_E
10	Shopping	0.67	0.82	0.65	0.82
	Nature & science	0.69	0.85	0.63	0.79
	Sports	0.67	0.84	0.65	0.81
	Business & politics	0.70	0.86	0.67	0.84
	Travel & leisure	0.67	0.83	0.65	0.80
	Entertainment	0.70	0.85	0.68	0.82
50	Shopping	0.84	0.94	0.73	0.89
	Nature & science	0.84	0.94	0.74	0.89
	Sports	0.86	0.94	0.77	0.90
	Business & politics	0.87	0.96	0.77	0.91
	Travel & leisure	0.83	0.93	0.73	0.88
	Entertainment	0.85	0.94	0.76	0.89

tributed, results across different subtopics for any topic. This can be due to the fact that Google and Bing give different priority or importance to different subtopics for a query which is observed here in terms of evenness.

Table 3.8 shows the average topic diversity of the search results from Google and Bing. As the richness was almost equal in both the Web search engines, then it is due to the evenness that Google has consistently higher diversity than Bing, in top-10 and top-50 Web search results, across all six categories. As in Google the diversity is consistently higher than Bing, and the reason for this is the higher evenness, which shows Google's approach for more balanced distribution of Web search results from different sub-topics linked to a main topic.

3.3 DIVERSITY OF ADVERT WEBSITES

3.3.1 Presentation and Analysis of Data

To analyse the sponsored search results diversity, we collected the adverts from the search results pages for all the queries in each category and then extracted the hostnames from the URL of each advert as we did earlier (see Section 3.1.2).

Table 3.9 shows the number of queries and adverts across all of the categories in Google and Bing. It can be seen that Bing has a much higher number

Table 3.8: Average topic diversity per query in Web search results.

		Google		Bing	
top-k	category	Simpson's D	Shannon's H	Simpson's D	Shannon's H
10	Shopping	5.21	0.72	5.10	0.73
	Nature & science	5.48	0.77	5.01	0.71
	Sports	5.26	0.75	5.05	0.72
	Business & politics	5.54	0.77	5.33	0.75
	Travel & leisure	5.24	0.74	5.08	0.72
	Entertainment	5.43	0.76	5.26	0.73
50	Shopping	6.61	0.83	5.76	0.79
	Nature & science	6.69	0.85	5.88	0.80
	Sports	6.74	0.84	6.06	0.81
	Business & politics	6.88	0.86	6.12	0.82
	Travel & leisure	6.50	0.83	5.77	0.78
	Entertainment	6.65	0.83	5.91	0.79

Table 3.9: The dataset of sponsored search results.

	Google	Bing
The number of all the queries	3516	3516
The number of queries with adverts	358	2822
The number of queries without adverts	3158	694
The number of adverts in all the queries	818	10275

of queries having adverts as well as the average number of adverts per query (for a query having advert) is higher in Bing than Google.

Table 3.10 lists the websites that occur most frequently in the sponsored search results from Google and Bing. Following the same pattern of organic search results, the distribution of websites in sponsored search results also exhibits a high skewness, for both of the Web search engines — a few big advertisers' websites occur very frequently while the majority of websites occur only a small number of times.

Fig. 3.9 shows the log-log plots of website frequency (against rank order) in the sponsored search results. Similar to what we see in the organic search results, it is apparent that the distribution of websites in the sponsored search results also roughly follows *Zipf's law* [64], though this time a better fit of the model could be found with the Zipf's law with exponential cut-off [22]. In other

Table 3.10: The websites that appear most often in the sponsored search results.

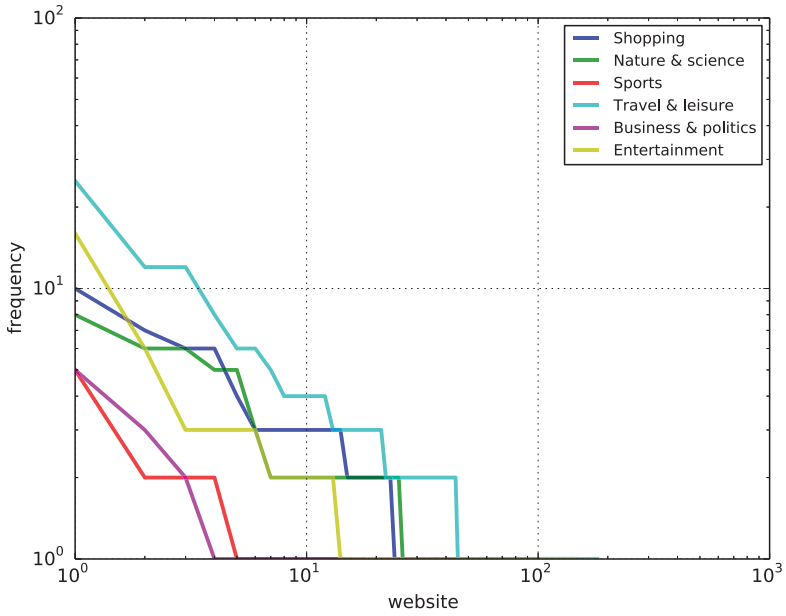
	Google		Bing	
rank	website	freq.	website	freq.
(1)	Amazon	62	Amazon	2291
(2)	wow	16	about	1092
(3)	tripadvisor	13	booking	249
(4)	audleytravel	13	shopzilla	207
(5)	autotrader	10	lowpriceshopper	171
(6)	booking	9	travelrepublic	147
(7)	ebay	8	televisionfanatic	112
(8)	marksandspencer	7	tripadvisor	109
(9)	carwow	7	viagogo	99
(10)	bluecross	6	lifescript	99

Table 3.11: The distribution of websites in the sponsored search results.

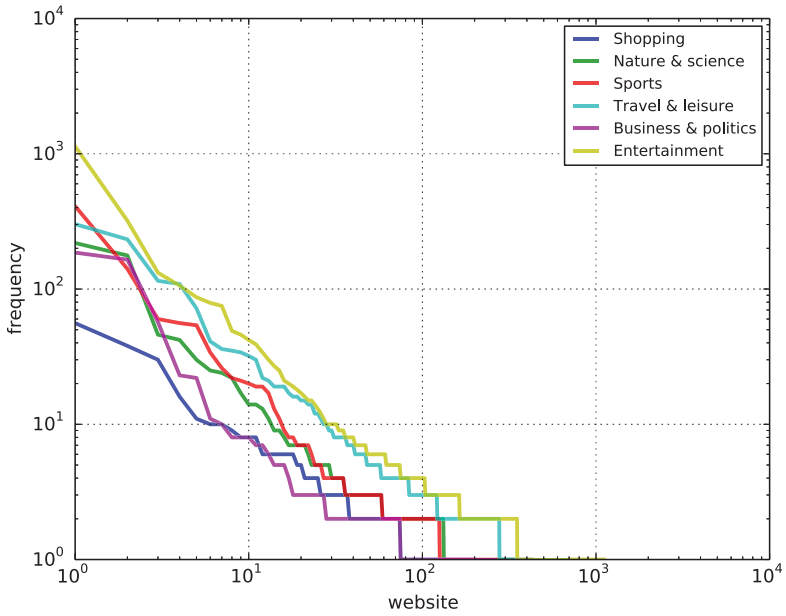
		Google			Bing		
top-k	category	#websites N	exponent s	cut-off q	#websites N	exponent s	cut-off q
50	Shopping	100	0.603	1.002	198	0.859	1.001
	Nature & science	82	0.556	1.001	424	1.122	1.002
	Sports	18	1.091	1.102	460	1.459	1.007
	Business & politics	50	0.823	1.038	259	0.000	0.601
	Travel & leisure	180	0.812	1.006	678	0.834	0.973
	Entertainment	145	0.986	1.018	1110	1.684	1.005

words, the frequency of the i-th popular website, f_i, should be proportional to q^i/i^s, where s is the exponent for the Zipfian distribution and q is the cut-off parameter.

Table 3.11 shows the number of distinct websites N together with the Zipfian exponents s as well as the cut-off parameter q for the sponsored search results in each category from Google and Bing, where q and s were estimated by fitting the log-log plot. Using Zipf's law with exponential cut-off, the coefficient of determination for the regression, R^2, is greater than 0.91 for all the categories, which confirms a very good fit of the model [57].



(a) Google



(b) Bing

Figure 3.9: The log-log plots of websites frequency in the sponsored search results.

3.3.2 Experiments and Results for Diversity in Advert Websites

Fig. 3.10 shows the website richness of the sponsored search results from Google and Bing (see also Table 3.11). It can be seen from this figure that in all the categories, the website richness of Bing's sponsored search results is much higher than Google's, i.e. Bing seems to use many more advertising

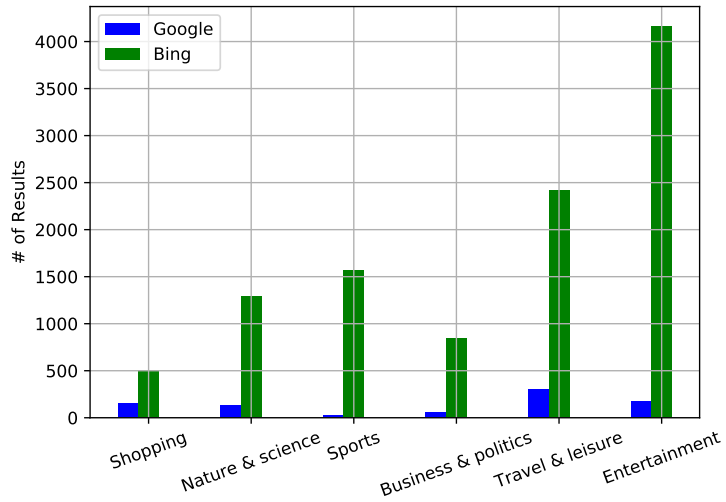


Figure 3.10: The websites richness of the sponsored search results.

Table 3.12: The websites evenness of the sponsored search results.

		Google		Bing	
top-k	category	Simpson's D_E	Shannon's H_E	Simpson's D_E	Shannon's H_E
50	Shopping	0.576	0.952	0.191	0.860
	Nature & science	0.634	0.957	0.049	0.752
	Sports	0.687	0.946	0.022	0.670
	Business & politics	0.762	0.974	0.011	0.221
	Travel & leisure	0.353	0.927	0.020	0.515
	Entertainment	0.427	0.945	0.006	0.543

Table 3.13: The websites diversity of the sponsored search results.

		Google		Bing	
top-k	category	Simpson's D	Shannon's H	Simpson's D	Shannon's H
50	Shopping	57.606	1.903	37.734	1.976
	Nature & science	51.98	1.831	20.933	1.977
	Sports	12.36	1.187	10.330	1.783
	Business & politics	38.101	1.655	2.951	0.533
	Travel & leisure	63.458	2.091	13.253	1.457
	Entertainment	61.965	2.043	6.387	1.655

sources than Google, which is not to be expected since Google has more traffic or users than Bing [1]. So apparently if advertisers show their adverts on Google, they are more likely to have more customers or users.

Table 3.12 shows the website evenness of the sponsored search results from

Google and Bing. Contrary to the website richness, in all the categories, Google has much higher website evenness in its sponsored search results than Bing, i.e. Google tends to spread the available advertising slots more evenly to different advertising sources than Bing does.

Regarding the website diversity in the sponsored search results, as shown in Table 3.13, Google consistently has higher Inverse Simpson's diversity than Bing in all the categories (following the same pattern as for website evenness), whereas using Shannon's diversity, the picture is not so clear.

Overall, what we can say for the sponsored search results is that Bing has higher richness while Google has higher evenness.

3.4 SIGNIFICANCE TESTS

To assess whether the diversity difference between Google and Bing, i.e. $|D^{(G)} - D^{(B)}|$ or $|H^{(G)} - H^{(B)}|$, is significant or not, we have performed a *nonparametric* statistical significance test [80]. It works as follows. For a given category, suppose that Google has m_1 search results and Bing has m_2 search results. We would first pool those $m_1 + m_2$ search results together and *randomly* partition the combined set of search results into two subsets of size m_1 and m_2 respectively, then calculate the diversity index of each subset, and finally check whether the simulated absolute diversity difference between those two random subsets is greater than the observed absolute diversity difference between Google and Bing. Iterating the above process for a large number of times (10,000 times in our experiment), the proportion of such random partitions with the simulated diversity difference greater than the observed diversity difference would provide the estimated p-value for this two-sided test. If the p-value is very small (say less than 0.01), we could confidently reject the null hypothesis that there is no real difference between Google and Bing on diversity.

Table 3.14 shows the statistical significance test outcomes for the comparison of Google and Bing in terms of the organic search results diversity see Section 3.1. For almost all of the categories, the p-value is far less than 0.01. The only exceptions are Simpson's D for the top-10 organic search results in category "Nature & Science" and Shannon's H for the top-50 organic search results in category "Business & Politics". Thus we can confirm that Google

Table 3.14: The statistical significance test results for comparing Google and Bing in terms of the organic search results diversity.

		Simpson's D		Shannon's H	
top-k	category	$\lvert D^{(G)} - D^{(B)} \rvert$	p-value	$\lvert H^{(G)} - H^{(B)} \rvert$	p-value
10	Shopping	22.399	0.0008	0.244	0.0000
	Nature & science	3.134	0.3744	0.116	0.0000
	Sports	8.600	0.0000	0.250	0.0000
	Business & politics	14.876	0.0000	0.349	0.0000
	Travel & leisure	8.514	0.0005	0.213	0.0000
	Entertainment	4.040	0.0000	0.242	0.0000
50	Shopping	76.606	0.0000	0.200	0.0000
	Nature & science	200.615	0.0000	0.334	0.0000
	Sports	67.283	0.0000	0.178	0.0000
	Business & politics	57.461	0.0000	0.040	0.5700
	Travel & leisure	136.389	0.0000	0.255	0.0000
	Entertainment	77.694	0.0000	0.172	0.0000

and Bing are indeed significantly different from each other, from the perspective of organic search results diversity. The same process can be applied to the sponsored search results diversity and topic diversities straightforwardly.

3.5 CONCLUSION

In this chapter first we have theoretically analysed how the diversity of Web search results is determined by the Zipfian distribution of websites. Notably, there is a tension between richness and evenness for a given Zipfian distribution, but richness matters more than evenness, because in Zipfian distribution when the zipfian exponent s is kept constant then the diversity increases with increasing richness see Section 3.1.3, so the overall diversity of search results would benefit from a longer tail of websites.

After that, we have empirically analysed how Google and Bing compare against each other over the diversity of their search results. Specifically, for organic search, Google is more diverse in the top-50 search results, while Bing is more diverse in the top-10 search results; for sponsored search, Google has higher evenness, while Bing has higher richness; for topic diversity, Google is

more diverse in top-10 and top-50 search results, while the richness is equal in both Google and Bing.

For the experiments and analysis in our thesis we have used the popular queries that are the result of most of the searches over a popular Web search engine (Google) and hence these are the queries which are considered important by WSEs so as to satisfy most of its users. Though the queries which are not popular can be important e.g. “breast cancer symptoms” or “fire emergency”. But in our thesis we have considered only the popular queries from Google top charts and the procedure could be followed to any set of queries from the users.

For the future work, as it is mentioned in Section 3.1.4, it would also be interesting to see the effect of diversity on the Web search results, when the reliability or credibility of the sources, websites in our case, is also considered.

Although in this chapter we focused on static diversity only. It would also be interesting to model how the diversity of Web search results changes with respect to time, as ecologists often do for species diversity [71], which is the topic of Chapter 5.

CHAPTER 4

PREDICTING QUERIES LIFETIME BY QUERIES COVERAGE DIVERSITY

As a Web search engine is limited by its coverage of the whole Web to fulfil the requirements of users from diverse backgrounds. Other than diversity in organic websites and topic diversity in Web search engines it is also important to have an insight into how the Web search engines compare in terms of coverage of the Web for popular queries.

It is also very helpful to see if there is any connection between diversity and the prediction of lifetime of popularity of a query.

In this chapter we shall first analyse the query coverage diversity in Section 4.1. Afterwards we show how the diversity related factors e.g. richness and evenness can be used to predict the lifetime of popularity of a query.

4.1 DIVERSITY OF QUERIES COVERAGE

4.1.1 Overview

It is interesting to see and compare Google and Bing for the total coverage of the Internet and how diverse it is for queries in the six different categories that we consider. For this purpose we use the *query results diversity*, in which we measure and compare the diversity of the total number of results in Google and Bing for most popular queries in the Google top charts[1].

[1] `http://www.google.com/trends/topcharts`

Table 4.1: The data set for the number of search results for queries from Google and Bing.

Category	# queries	Google # results (10^9)	Bing # results (10^9)
Shopping	99	12.13	0.64
Nature & science	340	55.77	18.93
Sports	709	20.98	5.51
Business & politics	375	31.37	3.24
Travel & leisure	528	154.00	14.39
Entertainment	1521	295.00	47.37

4.1.2 Presentation and Analysis of Data

In order to collect the data for the number of search results per query for our investigation, we first gathered all the popular queries (aka "hot searches"), as mentioned in Section 3.1.2 ,over 114 months from January 2004 until June 2013 in six representative categories of Google Top Charts[2]. The six categories are: (i) Shopping, (ii) Nature & Science, (iii) Sports, (iv) Business & Politics, (v) Travel & Leisure, (vi) Entertainment.

Thereafter we retrieve the figure for the total number of search results for every query from the Google and Bing search engines. Table 4.1 shows the number of search queries and the number of search results for each category in this large real-world data set. It can be seen here that Google has higher number of search results for queries than Bing, across all the categories.

Fig. 4.1 shows the average number of results per query across all the categories in Google and Bing. It is seen here that Google and Bing have the different average number of results per query in any category, i.e. Google has greater average number of results per query in Travel & leisure followed by Entertainment, Nature & science, Shopping, Business & politics and Sports category has the least number of results per query. On the other hand, in Bing has greater average number of results per query in Nature & science followed by Entertainment, Travel & leisure, Business & politics, Sports and Shopping category has the least number of results per query. The difference here could be because of how these two Web search engines update their index of Web

[2] `http://www.google.com/trends/topcharts`

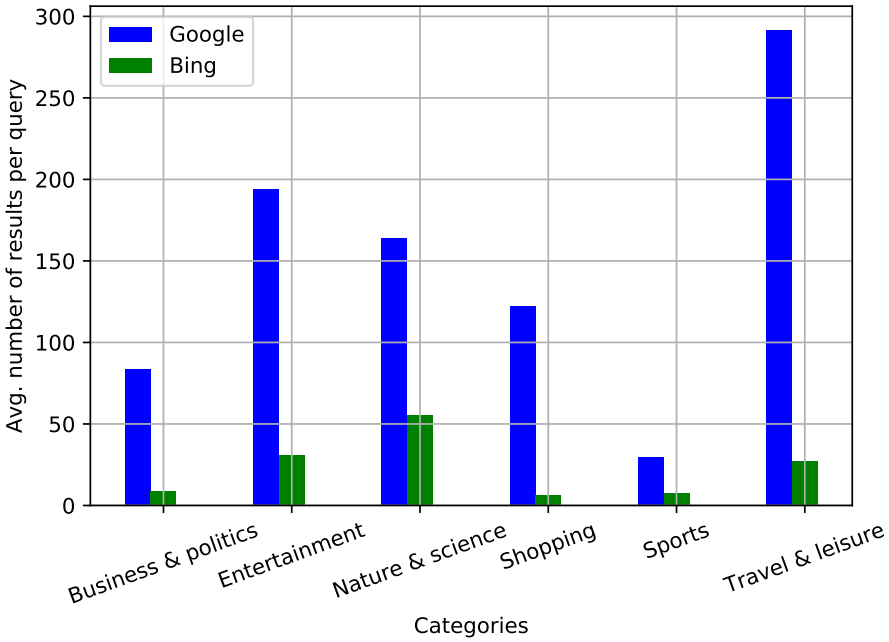


Figure 4.1: Average number of results per query, in millions (10^6).

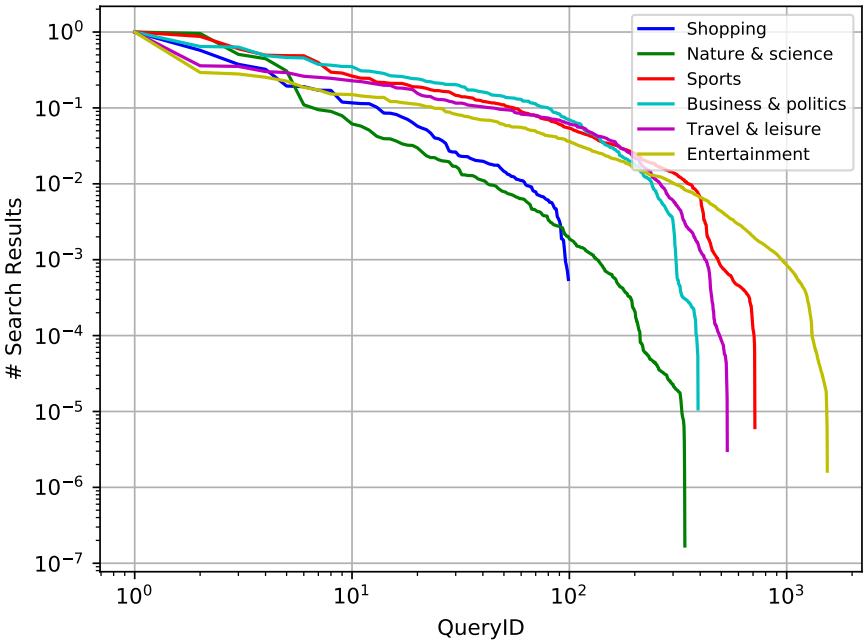
pages crawled from the Internet.

Fig. 4.2 shows the log-log plots of the total number of search results for queries in different categories for Google and Bing, respectively. Where the total number of results are normalised against maximum number of results for query. As the coefficient of determination, R^2, is more than 96% for the queries in all categories, we can say that it follows the Power law with exponential cutoff[22]. The frequency f_i, the total number of results, for the i-th query is directly proportional to q^i/i^s, where s is the exponent characterising the distribution and q is the cut-off parameter.

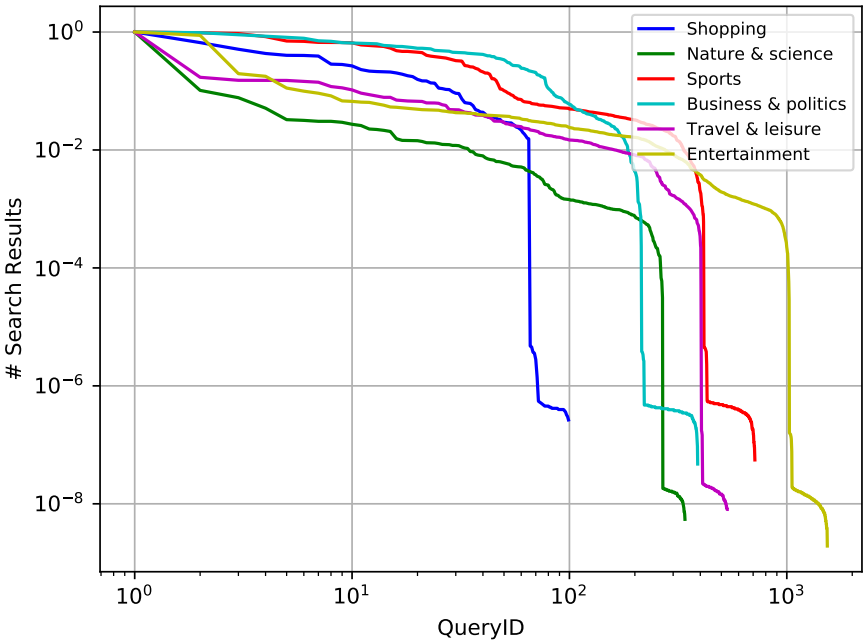
Without loss of generality, we could rank all the N queries according to their frequencies, the total number of results, then the proportion of the i-th query would be given by

$$p_i = \frac{f_i}{\sum_{k=1}^{N} f_k} = \frac{q^i/i^s}{\sum_{k=1}^{N}(q^k/k^s)} \,. \tag{4.1}$$

Table 4.2 shows the number of queries N together with the exponent s and cut-off parameter q, in each category from Google and Bing, where s and q were estimated by fitting the log-log plot through regression.



(a) Google



(b) Bing

Figure 4.2: The log-log plots of the total number of search results for queries.

Table 4.2: Query results distribution.

		Google		Bing	
Category	# queries	s	q	s	q
Shopping	99	0.738	0.976	0.257	0.961
Nature & science	340	1.153	0.998	0.000	0.619
Sports	709	0.557	0.996	0.000	0.598
Business & politics	375	0.428	0.991	0.000	0.312
Travel & leisure	528	0.517	0.996	0.822	0.910
Entertainment	1521	0.655	0.998	0.505	0.964

4.1.3 Experiments and Results for Diversity in Queries Coverage

Currently, Google and Bing are the two major Web search engines for English users. How do they compare against each other in terms of the diversity of their number of search results?

To calculate the diversity for queries coverage, we use Inverse Simpson's index and Shannon's diversity index as defined in Eq. (2.4) and Eq. (2.5) respectively. Where N is the total number of queries per category and p_i is the proportion of the total number of results for ith query with respect to the total number of results for all the queries per category.

Fig. 4.3 shows the queries coverage per category in Google and Bing (see also Table 4.1). An apparent pattern across all categories in this bar chart is that Google's coverage is consistently higher than the coverage in Bing.

Richness, number of queries per category, for Google and Bing is shown in Table 4.2. Which in this case remains exactly same in both Web search engines.

Table 4.3 shows Inverse Simpson's evenness D_E and Shannon's evenness H_E for the total number of search results in queries for Google and Bing. It is shown here that Google has higher Inverse Simpson's evenness and Shannon's evenness than Bing's in each category with the only exception of the *Shopping* category. Having more evenness in Google reflects the fact that Google has less differences in the total number of search results between queries for each category compared to Bing, whereas in the Shopping category it is vice versa.

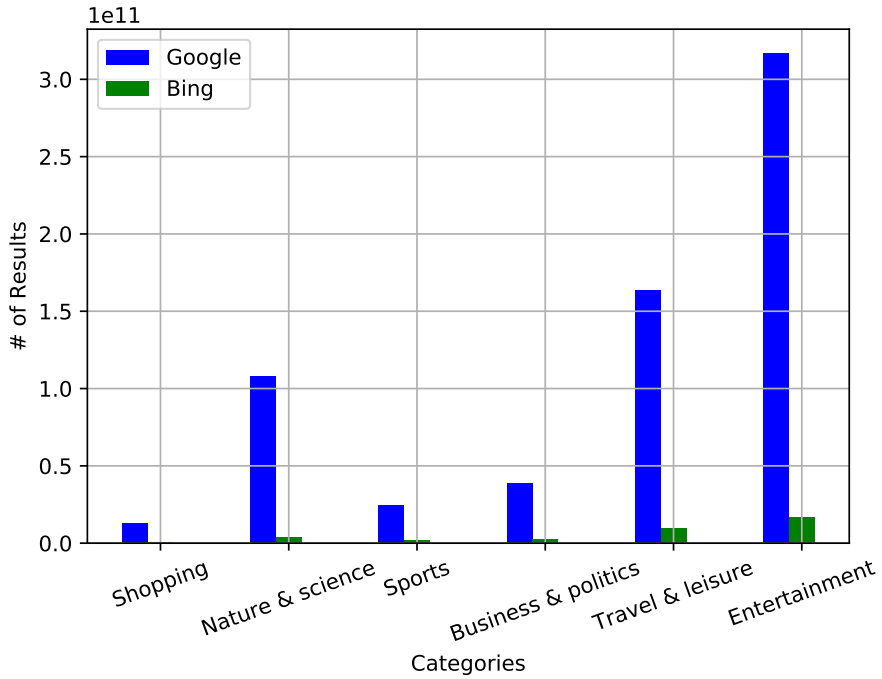


Figure 4.3: Query coverage results.

Table 4.3: Query results evenness.

	Google		Bing	
Category	Simpson's D_E	Shannon's H_E	Simpson's D_E	Shannon's H_E
Shopping	0.176	0.770	0.336	0.845
Nature & science	0.036	0.647	0.013	0.299
Sports	0.157	0.840	0.006	0.256
Business & politics	0.241	0.855	0.005	0.152
Travel & leisure	0.226	0.872	0.012	0.389
Entertainment	0.082	0.810	0.015	0.504

It can be seen from Table 4.4 that Google's diversity is consistently higher than Bing's in all categories with the only exception of the Shopping category — the same pattern that the evenness follows. It is rather very interesting to see the distribution of results is more balanced in Bing than Google in shopping category. This shows Bing's focus and interest in shopping category.

Table 4.4: Query results diversity

	Google		Bing	
Category	Simpson's D	Shannon's H	Simpson's D	Shannon's H
Shopping	17.427	1.537	33.310	1.686
Nature & science	12.255	1.636	4.253	0.758
Sports	111.350	2.394	3.979	0.728
Business & politics	90.297	2.200	1.908	0.392
Travel & leisure	119.388	2.375	6.561	1.060
Entertainment	125.293	2.577	22.945	1.605

4.2 PREDICTING THE LIFETIME OF QUERIES

Herein we analyse the factors which affect query popularity and investigate for any diversity-related factors influencing such a popularity.

The popularity of a query is determined by the number of months it remains in the Google top charts. Apparently, if a query remains more number of months in the Google top charts then it is more popular than the queries that remained for less number of months.

For the purpose of predicting the lifetime or popularity of a query, i.e the number of months it shall be in the Google top charts, we use Cox proportional hazard regression model [24]. The Cox proportional hazard regression model is a statistical model which is used in survival analysis. It deals with the survival time until an event of failure or death. It is used in areas, such as biology, sociology, economics and engineering among others [48]. The Cox proportional hazard regression model is defined as:

$$h(t) = h_0(t) \times exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) \; . \tag{4.2}$$

where $x_1, ..., x_k$ are the covariates that influence the survival duration, i.e. it affects the number of months for a query to remain in Google top charts. The

effects of these covariates are proportional as these are introduced through an exponential function. $\beta_1, ..., \beta_k$ are the coefficient that shows the importance of corresponding covariates. $h_0(t)$ represents the prediction time in the absence of any of the covariates.

We get some insight into the factors, i.e. covariates affecting the survival, i.e. the number of months, for a query to remain in Google top charts. We analyse different combinations of covariates for prediction of query popularity. The covariates that we use are: 1. The number of adverts for a query in the search results page, 2. The total number of results for a query, 3. Shannon's diversity of query coverage for the category of a query (see Table 4.4), 4. Simpson's diversity of query coverage for the category of a query, 5. Shannon's Evenness of query coverage for the category of a query (see Table 4.3), 6. Simpson's Evenness of query coverage for the category of a query and 7. a binary value for whether the search results for a query contain the popular website, e.g. Amazon, Twitter or Youtube. (Shown as covariates in Table 4.5)

We train the Cox proportional hazard regression model with a different variation of these covariates and compare the c, concordance, index [33] which estimates the probability of concordance between predicted and observed responses. More specifically it is the fraction of pairs in the data where the observation with higher survival time has the higher probability of survival predicted by the model. The value of concordance, c index, is between 0 and 1, where 0.5 is the expected result from random prediction, 1.0 is the perfect concordance, and 0.0 is perfect anti-concordance which can be multiplied with -1 to get the perfect concordance.

Table 4.5 shows the value of the c index after using various combinations of covariates. As we can see from this table, by using only diversity dependent covariates the value of the c index is 0.569, which is not better than the value, 0.617, seen by using the combination of popular websites for queries. But the highest value of 0.626 is obtained by using the combination of both; this indicates that diversity plays some role in predicting the lifetime of a query.

4.3 CONCLUSION

In this chapter, we compared Google and Bing in terms of diversity of queries coverage. First, we showed how the diversity of queries coverage is determined

Table 4.5: c index results for various covariates in Cox proportional hazard regression model

Covariates	c index	Covariates	c index	Covariates	c index
# Adverts # Results Shannon Diversity Simpson Diversity Shannon Evenness Simpson Evenness	0.569	Amazon Facebook DailyMail news.google imges.google Imdb TheGuardian Twitter Youtube	0.617	# Adverts # Results Shannon Diversity Simpson Diversity Shannon Evenness Simpson Evenness Amazon Facebook DailyMail news.google images.google Imdb TheGuardian Twitter Youtube	0.626

by Power law with exponential cutoff, in both Google and Bing. Then we showed that although richness, i.e. the number of queries, remains same in both Web search engines, the diversity of queries coverage in all the categories follows the same pattern as evenness which is higher in Google with the only exception of shopping category which is higher in Bing.

Afterwards by using Cox proportional hazard regression model we analysed the factors affecting the lifetime of a query. We observed that highest value of the c, concordance, index is obtained with the inclusion of diversity-related factors, i.e. diversity and evenness for the category of a particular query, with other factors which affect the lifetime of a query.

CHAPTER 5

ADDITIONAL DIMENSIONS OF DIVERSITY

Other than analysing source diversity in Web search engines we analyse the diversity in other aspects of Web datasets. Specifically, we investigate how the diversity changes with regards to time and we also want to analyse the diversity when the categories under observation are overlapping categories, i.e. an item of the dataset could be in more than one category. For example consider a movie which can be in more than one genre, e.g. "star wars" movie could be in "adventure" as well as "sci-fi" genres. The diversity that we have analysed until now is the static diversity and is only for the datasets which have hard categories. However, in real-world we have the datasets which are dynamically changing and the datasets which have overlapping categories.

In this chapter we first introduce another dataset of movies in Section 5.1, which we use for analysing diversity in overlapping categories and also the diversity with respect to time. Afterwards in Section 5.2 we propose a new method to investigate the diversity in a dataset having overlapping categories. Finally in Section 5.3 we analyse that how diversity is changing with respect to time and present a method to get the information from changing diversity, which is not directly visible directly from diversity values.

5.1 DATASET

For our purpose of analysing the diversity with regards to time and the diversity for datasets with overlapping categories, we use movies datasets from two different sources. The first dataset that we call the-numbers dataset which we extracted from the Web pages of a feature rich and structured Website [1]. The second dataset that we are using is from a very well known Website for movies which is Imdb[2] movies dataset. The the-number dataset contains total gross earned for any movie and the genres for movies and it has 1604 movies in 6 years time from 2009 to 2014. On the other hand, Imdb dataset mentions the principal people linked with a movie, which includes, actor, actress, director, producer and writer for each movie. The Imdb dataset that we use for our analyses, spans from 1917 to 2017.

5.2 DIVERSITY IN OVERLAPPING CATEGORIES

5.2.1 Diversity of Genres for Movies

For the diversity analysis in terms of overlapping categories, we compared the diversity of genres for movies across six years in the the-numbers dataset. It could be interpreted as the inverse of the probability that two movies randomly selected belong to the same genre.

To calculate the diversity we used Inverse Simpson's index and Shannon's diversity index as shown in Eq. (2.4) and Eq. (2.5) respectively, where N is the number of distinct genres for all the movies in a year and P_i is the proportion of genre i in all the movies.

As a movie could have more than one genre, we calculate the proportion of genres across movies with two separate methods. The first method (*Method I*) is defined in Eq. (5.1) where G_i is the total number of movies in genre i and N is the number of distinct genres. This method assumes a separate virtual movie for each genre of a single movie. For example, if a movie has three genres then we assume three separate movies for this case. This way all the

[1] `http://www.the-numbers.com/`

[2] `http://www.Imdb.com/`

movies shall have a single genre.

$$P_i = \frac{G_i}{\sum_{j=1}^{N} G_j}. \tag{5.1}$$

The second method (*Method II*) to calculate the proportions is defined in Eq. (5.2) where M is the number of movies, WM_j is the weight of movie j which is always 1 in this case, and GM_j is the number of genres for movie j. WG_i is the weight of genre i with respect to the number of genres for a movie, across all the movies. This method does not consider separate movies for a movie having multiple genres instead the separate genres are given a weight.

$$WG_i = \sum_{j=1}^{M} \begin{cases} \frac{WM_j}{GM_j} & \text{if movie } j \text{ has Genre } i \\ 0 & \text{otherwise} \end{cases}$$

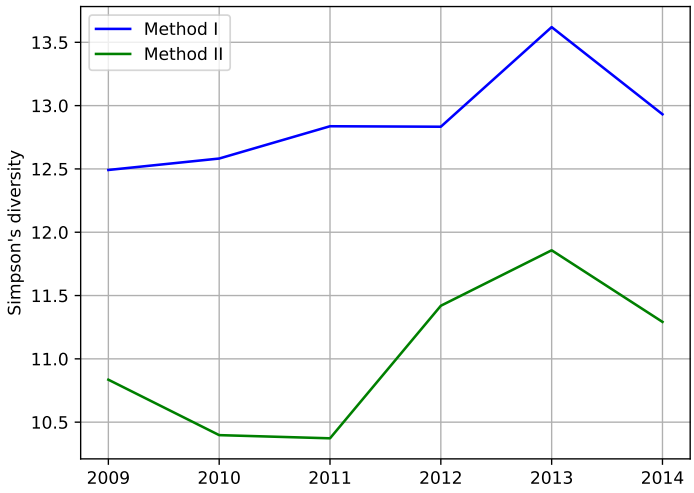
$$P_i = \frac{WG_i}{\sum_{j=1}^{N} WG_j}. \tag{5.2}$$

Fig. 5.1 shows the results for Inverse Simpson's diversity and Shannon's diversity, for the genres of movies across six years. This shows the diversity with two separate methods for the proportions of genres, i.e. *Method I* and *Method II*. It can be seen in the figure that the diversity results with *Method I* is always higher than the results with *Method II*. This is because of the reason that *Method I* assumes the maximum weight, i.e. 1, for each genre for a movie. Whereas, in *Method II* the weight is distributed across the genres for the same movie.

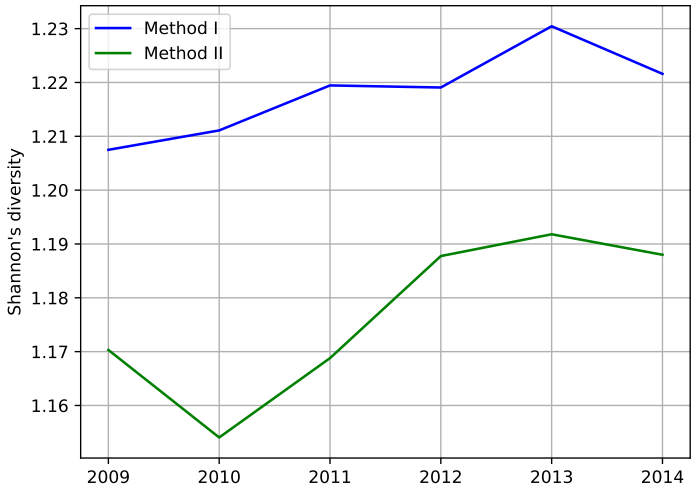
Fig. 5.2 shows the results of Evenness for Inverse Simpson's diversity and Shannon's diversity. It is observed here that as the richness, the number of genres, is constant across six years, i.e. 25 genres, the diversity follows the same pattern as evenness. As shown here, we can increase diversity in genres for movies with balancing the distribution of movies across all the genres of movies which shall also increase overall gross profit in movies as we show in Section 5.2.2.

5.2.2 Diversity of Gross Profit per Genre

In this section, we analysed the diversity in another dataset with the *Method I* and *Method II* mentioned in Section 5.2.1. We instigated the diversity of gross profit per genre.



(a) Simpson's diversity

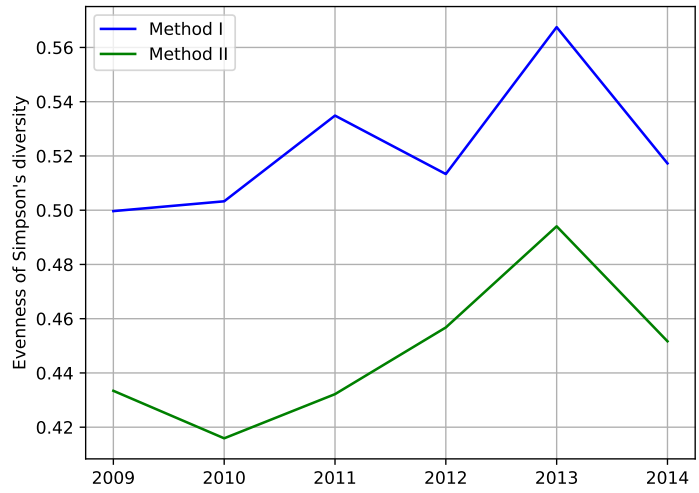


(b) Shannon's diversity

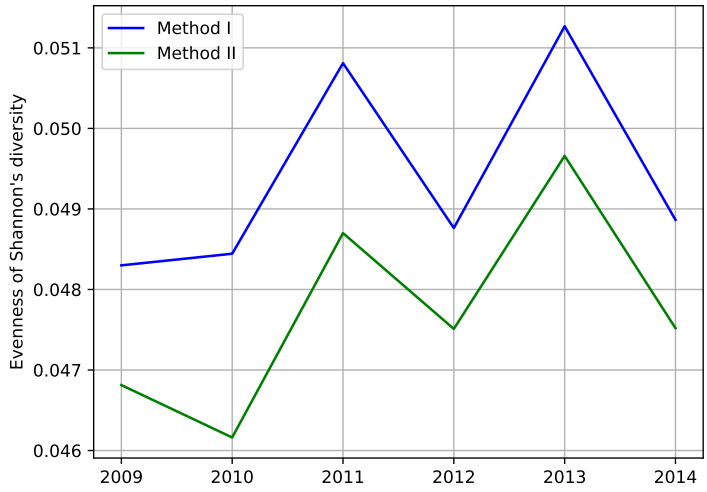
Figure 5.1: Genres *diversity* in movies with Method I and Method II.

For the first method, (*Method I*), in Eq. (5.1) we define G_i as the gross profit for all movies in genre i and N is the number of distinct genres. As mentioned earlier in Section 5.2.1, that this method assumes a separate virtual movie for each genre of a single movie. For example, if a movie has three genres then we assume its gross profit for all three separate genres for this case.

For the second method, (*Method II*), in Eq. (5.2) we define M as the number of movies, WM_j is the weight of movie j which is gross profit for movie j in this case, and GM_j is the number of genres for movie j. WG_i is



(a) Evenness for Inverse Simpson's diversity

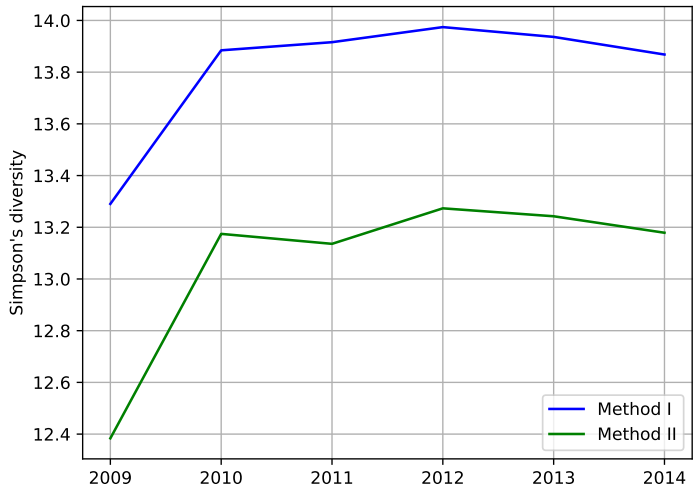


(b) Evenness for Shannon's diversity

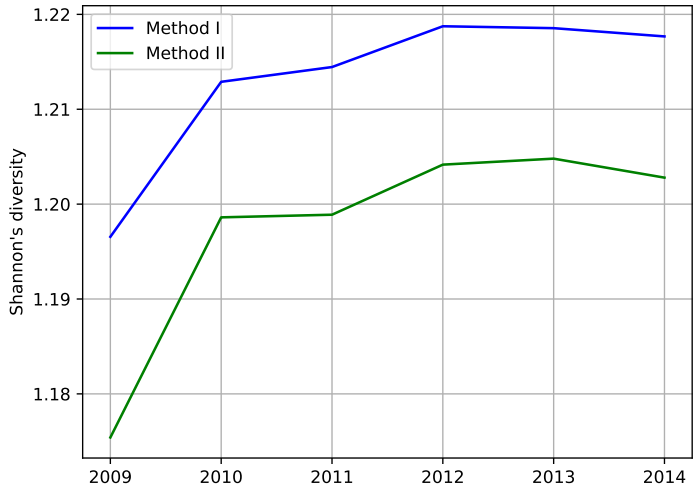
Figure 5.2: Genres *evenness* in movies with Method I and Method II.

the weight of genre i with respect to the gross profit.

Fig. 5.3 and Fig. 5.4 shows the results for diversity and evenness, respectively, of gross income per genre. As it was expected it follows the same pattern as the diversity of genres because apparently, the gross income per genre is dependant on the number of movies per genre.



(a) Simpson's diversity

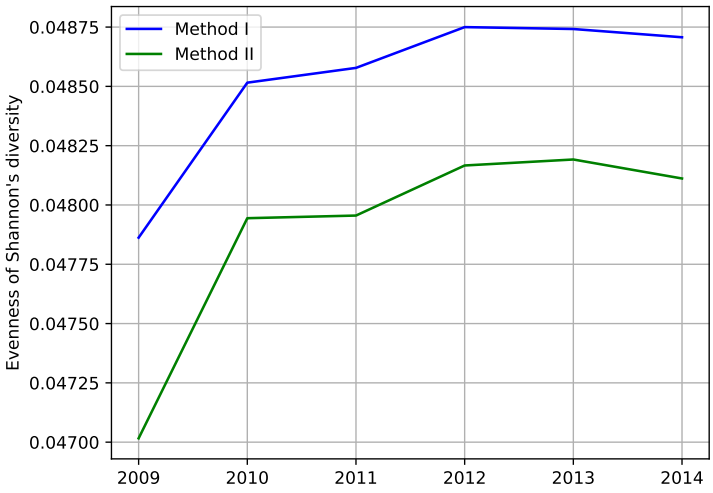


(b) Shannon's diversity

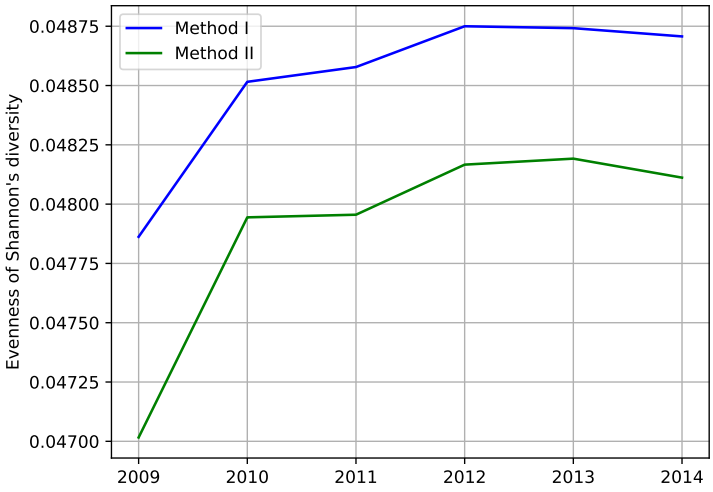
Figure 5.3: Gross income *diversity* in movies with Method I and Method II.

5.3 DIVERSITY WITH RESPECT TO TIME

For the diversity analysis with respect to time, we compared the diversity for five separate movie principal types, i.e. actor, actress, director, producer and writer from the year 1917 to 2017. It is shown in Fig. 5.5 that the number of movies listed at Imdb, for every principal type per year is different. It can be seen that the number of movies has grown substantially over the past 20 years. Production of more movies from past 20 years could be attributed to



(a) Evenness for Inverse Simpson's diversity



(b) Evenness for Shannon's diversity

Figure 5.4: Gross income *evenness* in movies with Method I and Method II.

creating more demand by reaching more public through advertisements, over the Web, which became possible after the increasing usage of the Internet over these years.

Fig. 5.6 shows the richness, which is the unique number of principals, per year for movies. The richness of all the principal types follows the same pattern as the number of movies produced across the years. Which also started increasing sufficiently around the same period, i.e. past 20 years.

To calculate the diversity we used Inverse Simpson's index as shown in

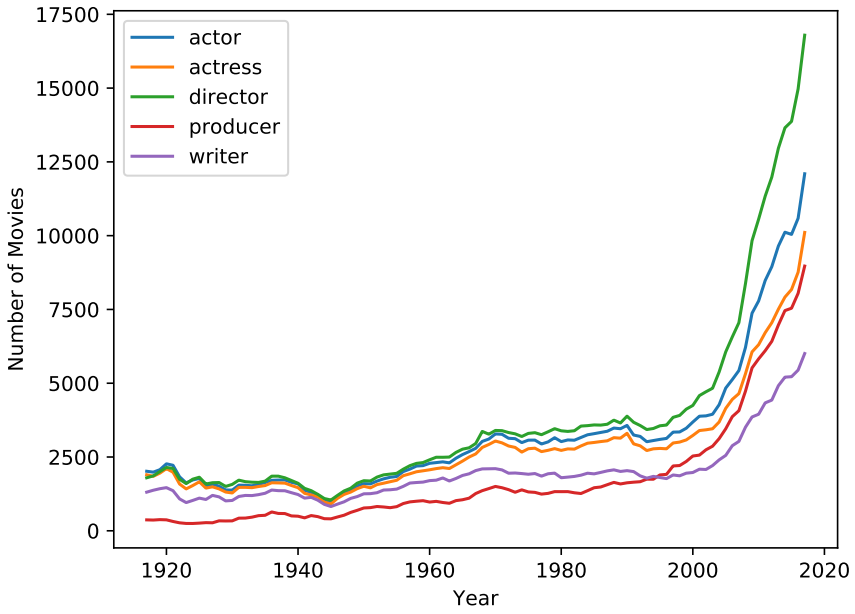


Figure 5.5: Number of movies per year in Imdb.

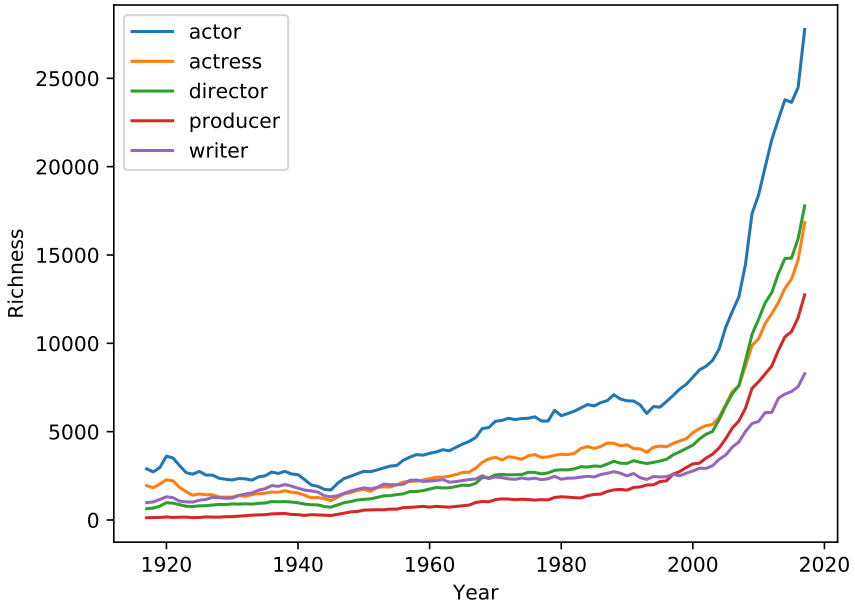


Figure 5.6: Richness per year in Imdb.

Eq. (2.4), where N is the different actors for the principal type actor and different directors for principal type director and so on, and P_i is the proportion of actor and director for the corresponding principal type. Fig. 5.7 shows the Inverse Simpson's diversity per year for different principal types in movies. It

can be seen here that diversity follows the same pattern as richness for every principal type. This increase in diversity can be attributed to more movies being produced every year as it is evident in the richness of movies. The high increase in number of movies produced from last 20 years can be because of increasing interest in people (consumers) by more advertisement and reaching more people through the invent of Internet.

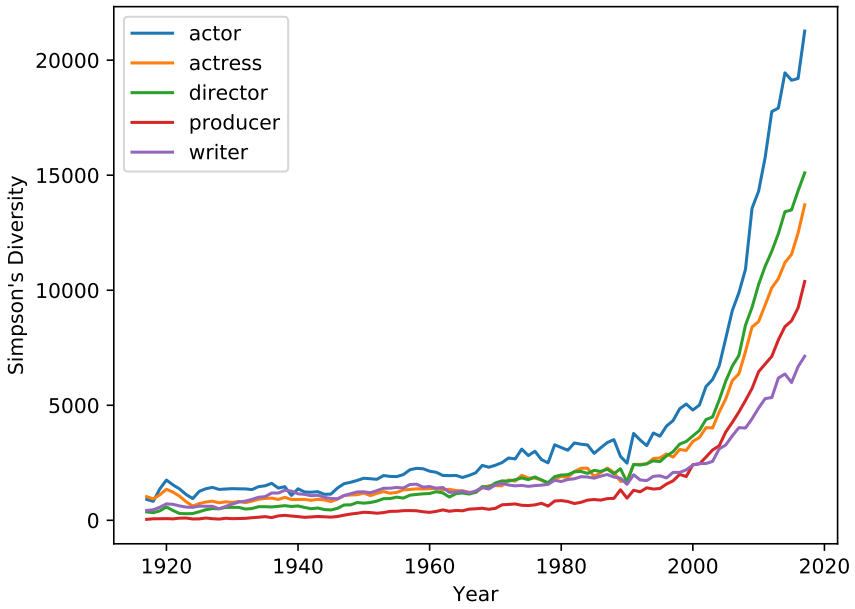


Figure 5.7: Inverse Simpson's Diversity per year in Imdb.

Fig. 5.8 shows the evenness per year. The pattern of evenness for all the principal types in a movie, across the years, is not clear from the figure. It has quite a few peaks and drops across the years.

To observe any trend and seasonality in the diversity and evenness across the years we used the time series analysis. For the time series analysis, we used only the principal type actor as all other principal types follow the same pattern. Fig. 5.9 shows the decomposed component plots, with the frequency of every 2 years, for the Inverse Simpson's diversity of principal type actor. It is shown here that the trend was quite apparent in the original data for Inverse Simpson's diversity. It is interesting to see the seasonality component which is varying from 10 to -10, in Inverse Simpson's diversity value, every two years.

Fig. 5.10 shows the decomposed component plots, with the frequency of

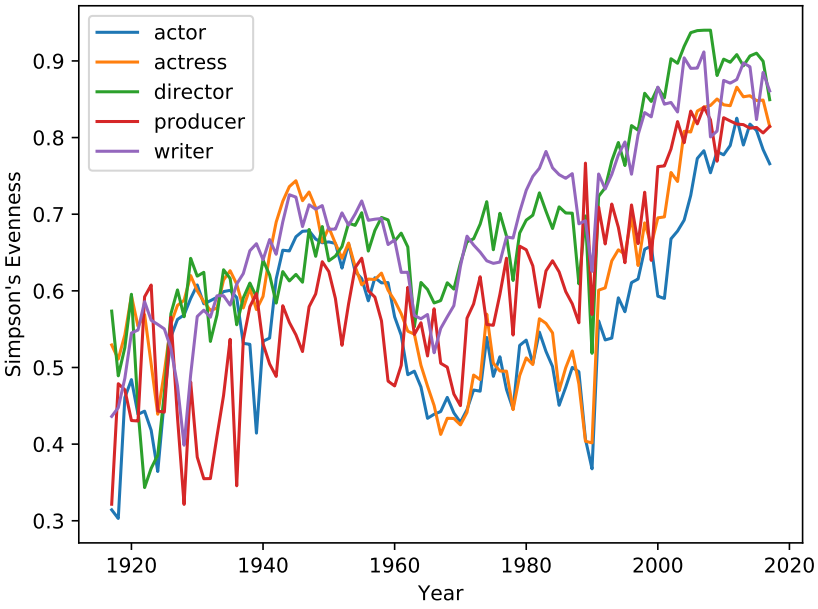


Figure 5.8: Inverse Simpson's Evenness per year in Imdb.

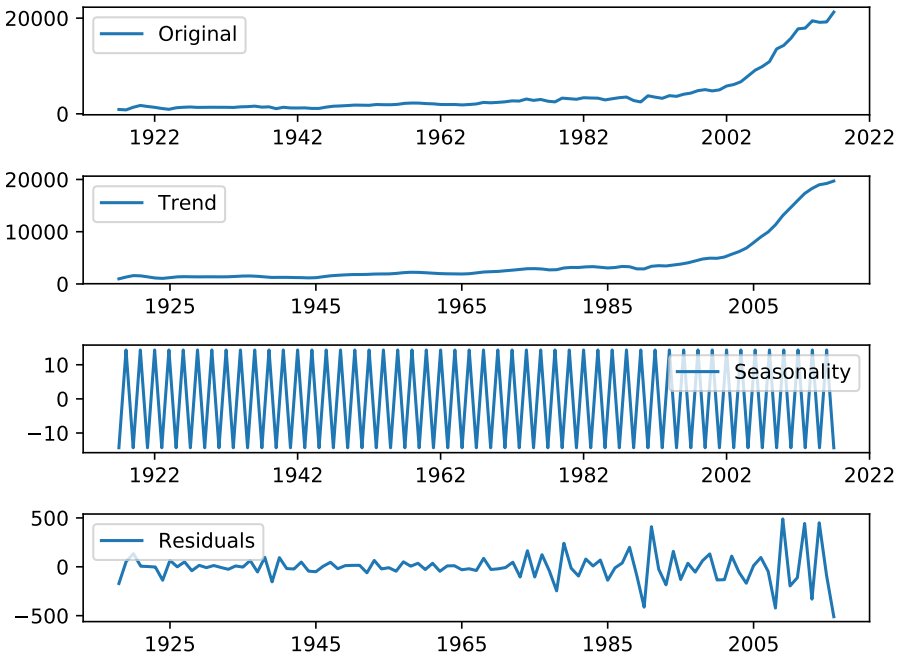


Figure 5.9: Time series decomposed component plots of Inverse Simpson's diversity for actors.

every 25 years, for the Inverse Simpson's diversity evenness of principal type actor. It is interesting to see the pattern and the seasonality which was not very obvious in the original data of evenness for Inverse Simpson's diversity.

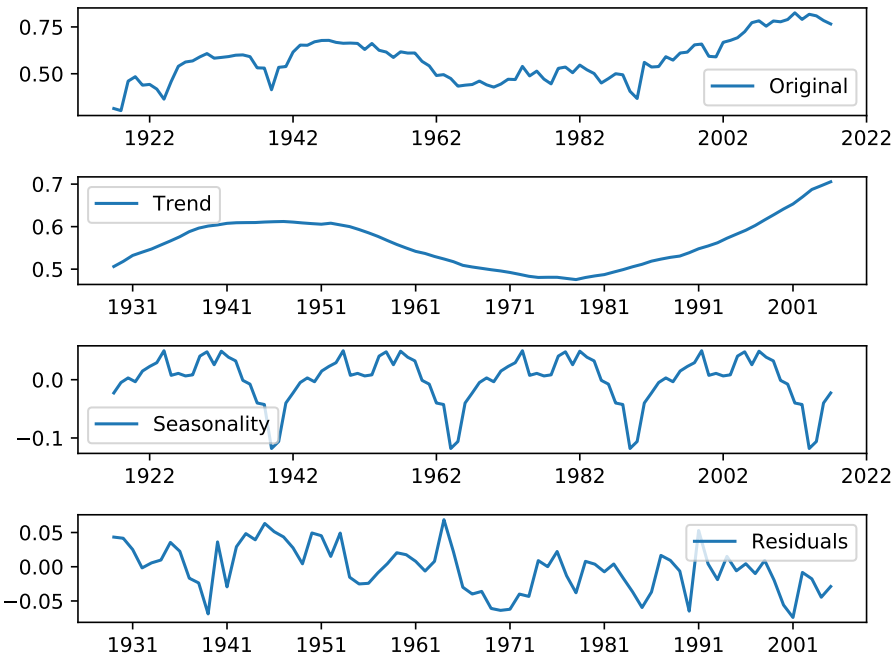


Figure 5.10: Time series decomposed component plots of Evenness for Inverse Simpson's diversity for actors.

5.4 CONCLUSION

In this chapter first, we showed how the diversity could be analysed with two different methods for calculating the proportions of categories which are overlapping, i.e. the categories are not hard and an individual item could be in more than one category. Afterwards, we analysed how the diversity changes with respect to time. Notably, it is shown that how we can use time series decomposed components to see the trend and seasonality which are not very obvious by just looking at the diversity measures.

For the future work, it would also be interesting to look into another dimension of diversity into Web search results, which is to have a look into how diversity changes when queries from multiple languages are considered. For example sometimes their is more ambiguity one language than another language e.g. a sentence in English language is usually more ambiguous than a sentence in French language.

CHAPTER 6

TWO APPLICATIONS OF DIVERSITY

In this chapter we present two real world applications of diversity in which we use diversity method to get some useful information. In Section 6.1 we show how we can estimate the number clusters from a dataset. We propose a new diversity method to estimate the number of balanced clusters and avoid any outliers. In Section 6.2 We show how the evaluation methods for diversified search, which are popular in the literature, can detect the changing levels of richness, evenness and relevance in Web search results.

6.1 ESTIMATING THE NUMBER OF CLUSTERS USING DIVERSITY

It is an important and challenging problem in unsupervised learning to estimate the number of clusters in a dataset. Knowing the number of clusters is a prerequisite for many commonly used clustering algorithms such as k-means. Here, we propose a novel *diversity* based approach to this problem. Specifically, we show that the difference between the global diversity of clusters and the sum of each cluster's local diversity of their members can be used as an effective indicator of the optimality of the number of clusters, where the diversity is measured by Rao's quadratic entropy. A notable advantage of our proposed method is that it encourages *balanced* clustering by taking into account both the sizes of clusters and the distances between clusters. In other words, it is less prone to very small "outlier" clusters than existing meth-

ods. Our extensive experiments on both synthetic and real-world datasets (with known ground-truth clustering) have demonstrated that our proposed method is robust to clusters of different sizes and shapes, and it is more accurate than existing methods (including elbow, Caliński-Harabasz, silhouette, and gap-statistic) in terms of finding out the true number of clusters.

6.1.1 Overview

Clustering is an important unsupervised learning task aiming to group a collection of items into subsets (clusters) such that those within the same cluster are more closely related (similar) to each other than to those in different clusters [34]. For many commonly used clustering algorithms (such as k-means [34], k-medoids [34], Gaussian mixtures [92], and spectral clustering [59]), it is necessary to specify beforehand the number of clusters, a parameter often labelled k as in the k-means/k-medoids algorithm, to run the algorithm. However, we often do not have prior knowledge about the correct choice of k, and it is a very challenging problem to accurately estimate it by analysing the dataset itself only [54, 73, 39]. On one hand, increasing k will reduce the amount of error (in terms of data recovery [56]) in the resulting clustering, to the extreme case of full accuracy when $k = n$ the total number of items in the dataset. On the other hand, decreasing k will offer a higher compression ratio, to the extreme case of maximum compression when $k = 1$. The optimal choice of k probably lies somewhere in the middle ground, depending on the characteristics of the dataset such as its size, variance, and shape.

We propose a novel *diversity* based approach to the problem of estimating the number of clusters in a dataset. A notable advantage of our proposed method is that it encourages *balanced* clustering by taking into account both the sizes of clusters and the distances between clusters. In other words, it is less prone to "outlier" clusters (that are much smaller than most other clusters in the dataset) than existing methods. Such a property of clustering is usually desirable in practice. For example, when using a clustering algorithm to perform image segmentation [23], a very small cluster (segment) usually corresponds not to a complete meaningful object but only part of it, and therefore should be avoided. For another example, when using a clustering algorithm

to perform market segmentation [91], a very small cluster (segment) probably means that the market segment has too few customers to be profitable, and therefore should be discouraged. Obviously in some scenarios, small outlier clusters can be useful, e.g., for revealing exceptions or abnormalities in the data. However, there are many real-world applications where balanced clusters are preferred, which is the focus of this paper.

6.1.2 Related Work for Estimating the Number of Clusters

The problem of estimating the number of clusters k in a dataset has been studied extensively, and a number of methods have been proposed by researchers from various disciplines. In this section, we review a few representative ones.

The elbow method

The elbow method [83] examines the percentage of variance explained by the clustering as a function of the number of clusters k. If we plot the percentage of variance explained against k, the first clusters will be able to explain a lot of variance, but at some point the marginal gain will drop, giving an "elbow" in the graph. The optimal k is chosen at this point, as introducing more clusters would not give a better explanation of variance in the dataset, though such an "elbow" cannot always be unambiguously identified [44]. In this paper, we use a slight variation of this method which plots the curve of the intra-cluster variance [32]:

$$E(k) = \sum_{r=1}^{k} W(C_r) \ , \tag{6.1}$$

where $W(C_r)$ is the variance within the r-th cluster C_r.

The Caliński-Harabasz method

Milligan et.al. [54] compared 30 different methods to finding the estimated number of clusters in the dataset and found that the best performing method is given by Caliński and Harabasz [13]:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \ , \tag{6.2}$$

where $B(k)$ is the inter-cluster variance (i.e. the sum of squared distances for the k clusters), and $W(k)$ is the intra-cluster variance. Maximising $CH(k)$ against different values of k gives the estimated number of clusters.

The silhouette method

Rousseeuw et.al. [72] proposed the silhouette method, of which the main goal is to examine whether an item i is classified well in the cluster or not. For every item or point i, its silhouette is calculated as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \ , \tag{6.3}$$

where $a(i)$ is the average distance of item i to all the items in the same cluster and $b(i)$ is its average distance to all the items in the nearest cluster. The i-th item is well clustered if the value of $S(i)$ approaches the maximum which is 1; and a $S(i)$ value 0 means $a(i) = b(i)$, whereas $S(i)$ value -1 means item i belongs to the other cluster. After plotting the average $S(i)$ for all the items against different values of k from 1 to n, the maximum value of average $S(i)$ for all the items gives the estimated number of clusters, k, for the dataset.

The gap-statistic method

Tibshirani et.al. [84] proposed another method, gap-statistic, which compares intra-cluster variance with the expected values under the null reference distribution of the dataset. After clustering the dataset for different values of k, we get the intra-cluster variance for the observed and reference datasets and calculate the gap-statistic as:

$$Gap_n(k) = E_n^*\{\log(W(k))\} - \log(W(k)) \ , \tag{6.4}$$

where $W(k)$ is the total intra-cluster variance and E_n^* denotes the expectation under a sample of size n from the reference distribution. The gap-statistic measures the deviation of the observed $W(k)$ value from its expected value under the null distribution.

6.1.3 Our Approach for Estimating the Number of Clusters

One drawback of the above mentioned methods for estimating the number of clusters is that they could lead to very imbalanced clustering, where some "outlier" clusters are much smaller than the other clusters. This is often undesirable for real-life clustering applications (see Section 6.1.1). Here we propose a novel *diversity* based approach to the problem of estimating the number of clusters, which is less tolerant to such "outlier" clusters and encourages

balanced clustering by taking into account both the sizes of clusters and the distances between clusters.

The Diversity Method for Estimating the Number of Clusters

The requirement of balance among clusters, in fact, implies that there should be no particular cluster dominating the dataset, i.e. there should be a certain level of diversity among clusters.

To find out the true number of clusters in a dataset with n items, we use the output of the given clustering algorithm (such as k-means) and then measure the difference between the global diversity of clusters and the sum of each cluster's local diversity of their members, denoted by $Q(k)$ and given by

$$Q(k) = Div^G - \sum_{r=1}^{k} Div_r^L \ , \tag{6.5}$$

where Div^G is the global diversity of k clusters (with each cluster as a species) while Div_r^L is the local diversity of the r-th cluster (with each member item of the cluster as a species) as measured by Rao's quadratic entropy given in Eq. (2.6). We choose to use Rao's quadratic entropy [69] to measure the diversity of data, because it takes into account both the sizes of species (clusters) and the distances between species (clusters). We calculate the diversity based statistic $Q(k)$ for various values of k, i.e. for $k = 1$ to n, and the maximum value of $Q(k)$ should be able to tell us the true number of clusters in the dataset, i.e.

$$\widehat{k} = \arg\max_{1 \le k \le n} Q(k) \ . \tag{6.6}$$

The underlying intuition of this diversity method is that in a good clustering, the items within each cluster should be as homogeneous as possible, (i.e. less local diversity), while the clusters themselves should be as heterogeneous as possible, (i.e. more global diversity). The balance of cluster sizes is actually implied by a high level of diversity among clusters.

The approaches to estimating the number of clusters can be divided into two categories, global methods and local methods, as pointed out by Gordon [31]. The former evaluate some measure over the entire dataset and optimise it as a function of the number of clusters; the latter consider individual pairs of clusters and decide whether they should be amalgamated [84]. Obviously, the diversity method proposed by us is a global method. According

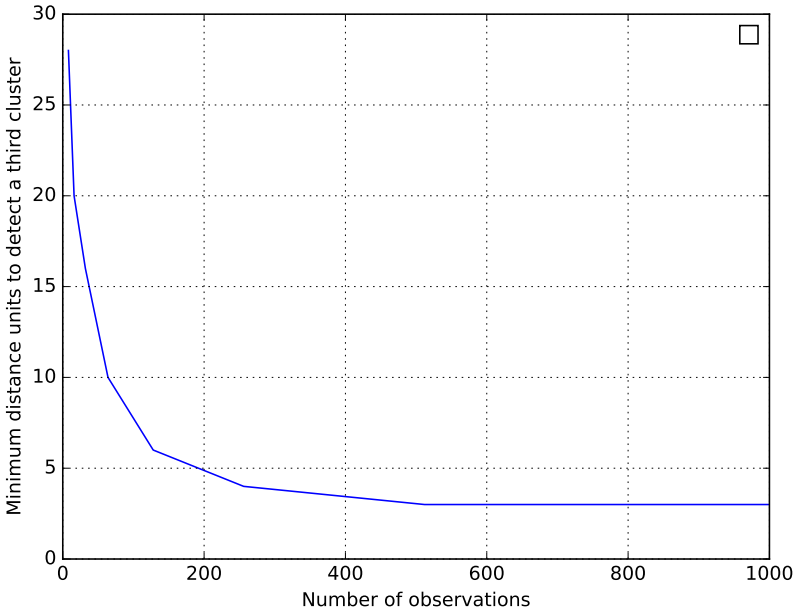


Figure 6.1: The trade-off between the sizes of clusters and the distances between clusters.

to Gordon [31], most global methods suffer from a serious disadvantage that they are undefined for one cluster, (i.e. $k = 1$) and therefore cannot be used to determine whether the dataset should be clustered at all. It is worth mentioning that our diversity method does not have this shortcoming: $Q(k)$ is well defined for $k = 1$, as we show later in Section 6.1.4.

6.1.4 Experiments for Estimating the Number of Clusters

Balance Between Size and Distance in Clusters

As can be seen in Eq. (2.6), Rao's quadratic entropy takes into account the sizes of clusters and the distances between clusters, which is important to achieve *balanced* clustering that is desirable in many real-life clustering applications.

For the purpose of investigating the trade-off between the sizes of clusters and the distances between clusters, we first create two clusters from two 2-dimensional standard normal distributions which have 1000 items each and are centred at $(0, 0)$ and $(0, 5)$ respectively, and then we create another cluster from one 2-dimensional standard normal distribution with varying number of items from 1 to 1000, (i.e. we obtain 1000 different datasets). Following this, we move the third cluster's centre (x, y) as follows: we keep y at 2.5 (halfway

from the first cluster's centre to the second cluster's centre), and gradually increase x from 0 to $+\infty$ until the third cluster is detected by our proposed diversity method as a separate, third, cluster.

The results of the simulation study are shown in Fig. 6.1, which indicate that using the diversity method to estimate the number of clusters, a small cluster needs to be distant from the other clusters in the dataset to be regarded as a separate cluster, otherwise it will be assimilated into another nearby cluster: the smaller the cluster, the larger its distance to the other clusters should be. In other words, the diversity method tends to avoid suggesting very small clusters unless they are very far away from the rest of the data.

Robustness of Our Method for Estimating the Number of Clusters

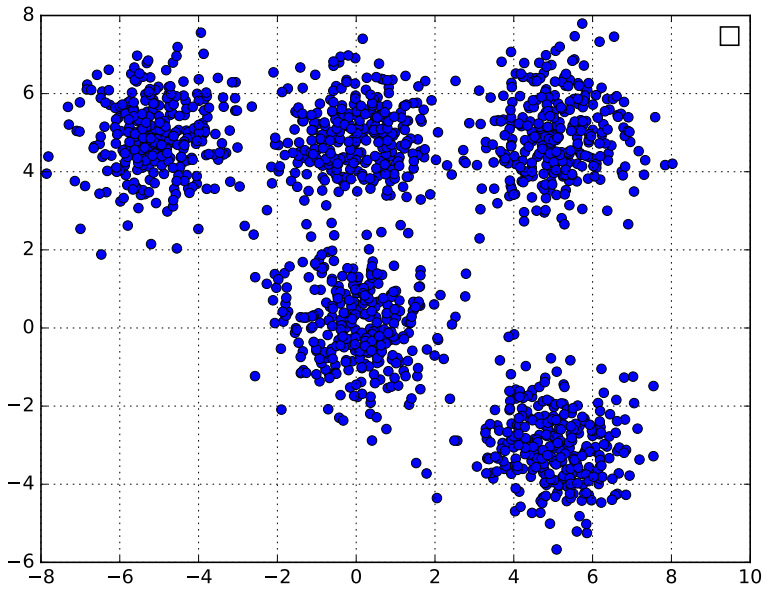
In this section, we investigate how robust our proposed diversity method is when it is applied to different types of datasets.

For this purpose, we create five synthetic datasets of different sizes, variances, and shapes. In addition, we also make use of three real-world datasets — Wine, Breast Cancer, and Thyroid Disease — from the UCI Machine Learning Repository [50]. On these synthetic and real-world datasets, we cluster the data points into k clusters with k from 1 to n (using k-means for the first three synthetic datasets and the first real-world dataset, but average-link hierarchical agglomerative clustering [53] for the remaining datasets), and calculate the value of $Q(k)$ for each k. The actual number of clusters in the dataset is estimated to be the k that maximises $Q(k)$ (see Section 6.1.3). It can be seen from the experimental results in Figs. 6.2 to 6.6, for both synthetic and real-world data, no matter what size, variance, or shape the dataset has, our proposed diversity method can successfully discover the correct number of clusters.

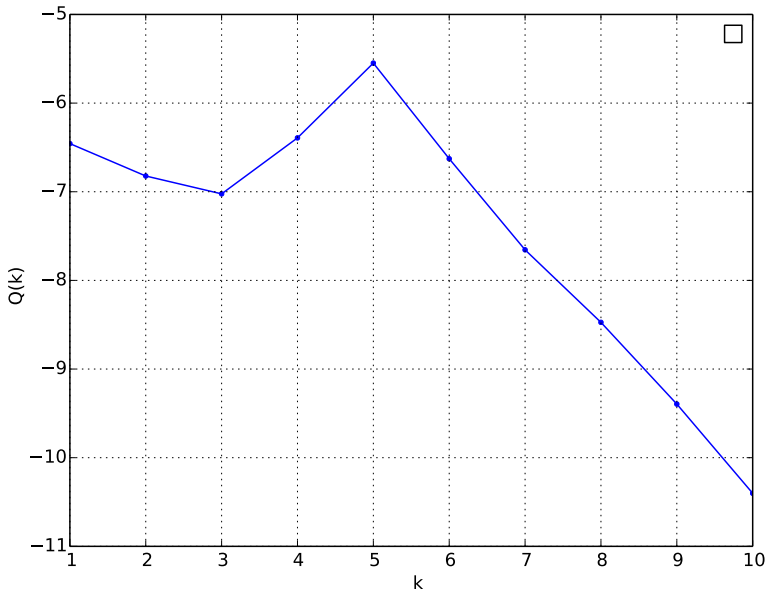
Comparison with Other Methods for Estimating the Number of Clusters

We use four synthetic datasets to evaluate and compare $Q(k)$ method to the other methods introduced in Chapter 2, i.e. elbow, Caliński-Harabasz, silhouette, and gap-statistic.

All those datasets differ in terms of the number of clusters, the number of dimensions, and the number of items. They are defined as follows.



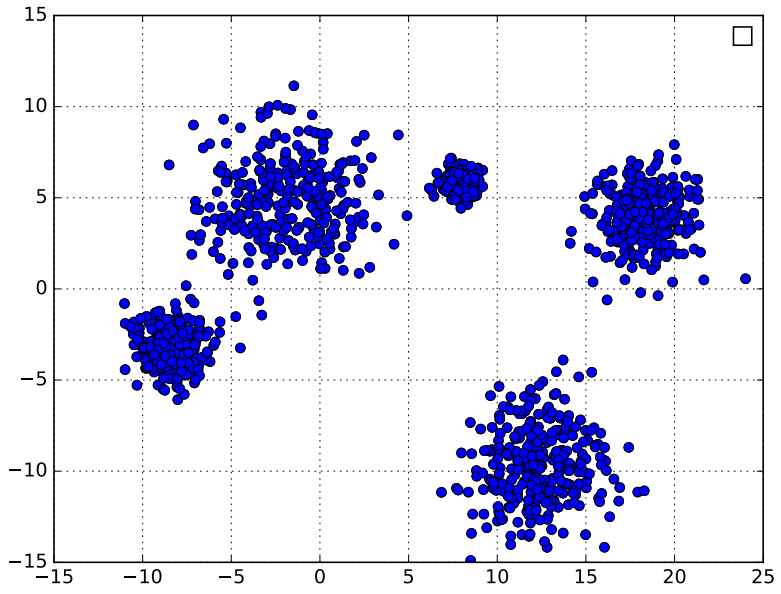
(a) Dataset



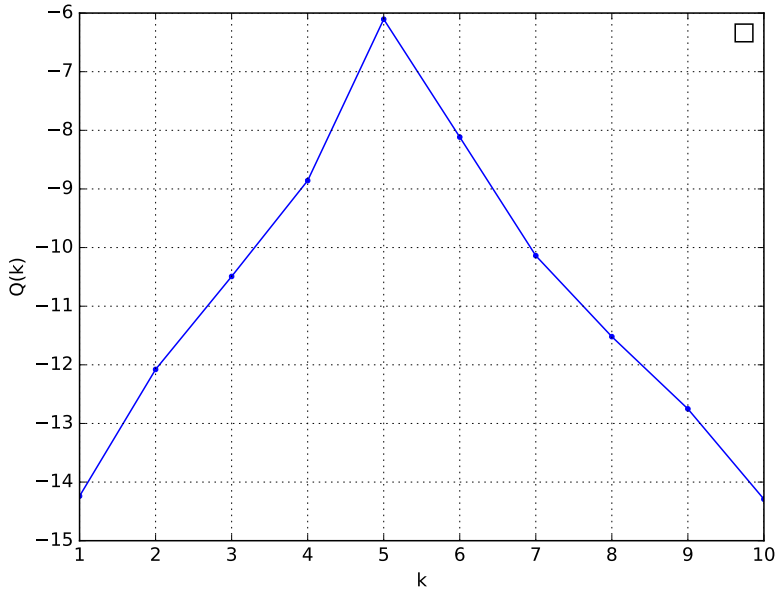
(b) Diversity based statistic $Q(k)$

Figure 6.2: Experiments on the synthetic dataset with five clusters with equivalent sizes and equivalent variances.

- (a) Four clusters in 2 dimensions; their sizes are 250, 250, 250, and 500 respectively; their centres are $(1, 3)$, $(0, 8)$, $(8, 0)$ and $(4, -2)$ respectively.
- (b) Four "normal" clusters and one small "outlier" cluster in 2 dimensions; the sizes of those "normal" clusters are 1000, 900, 900, and 850 respectively while the size of that "outlier" cluster is randomly set to a number between 50 and 100; the centres for all the clusters are chosen



(a) Dataset

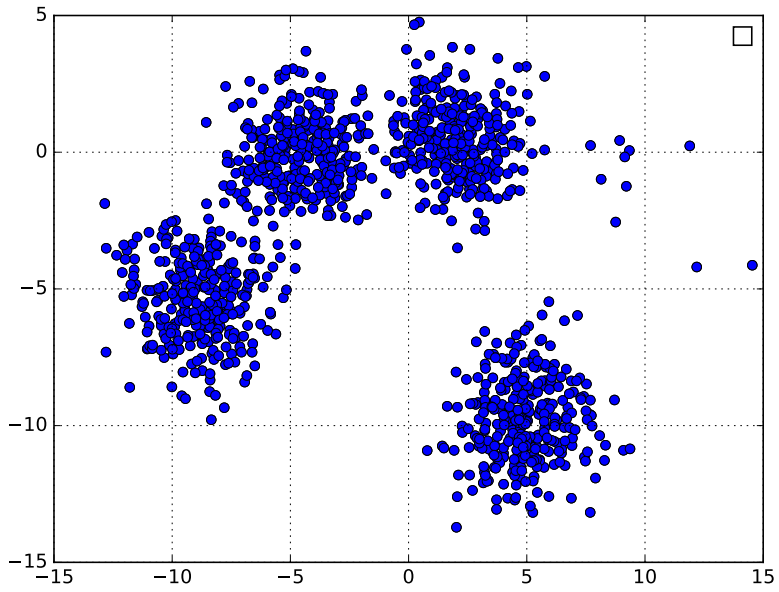


(b) Diversity based statistic $Q(k)$

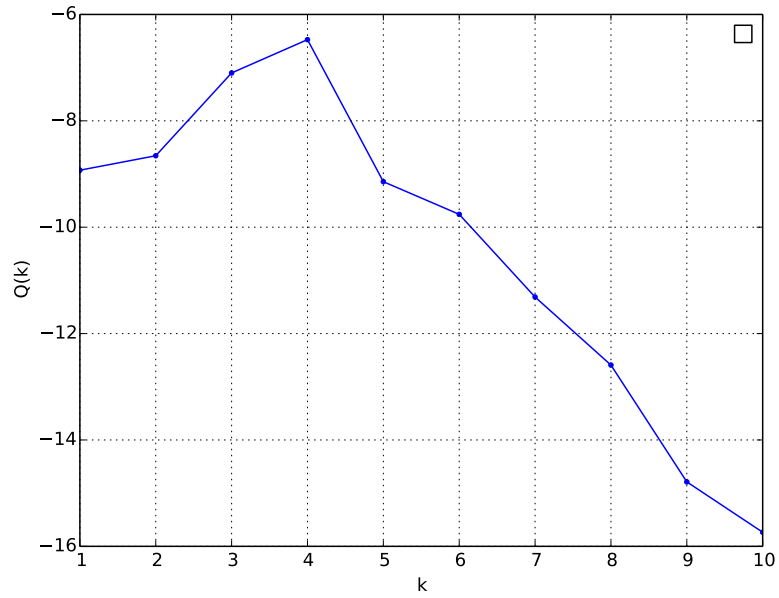
Figure 6.3: Experiments on the synthetic dataset with five clusters with equivalent sizes but different variances.

randomly.

- (c) Five clusters in 10 dimensions; their number of items are randomly set to either 50 or 100; their centres are chosen randomly.
- (d) Six clusters with the same settings as in 5 clusters except that the number of dimensions is 4.



(a) Dataset

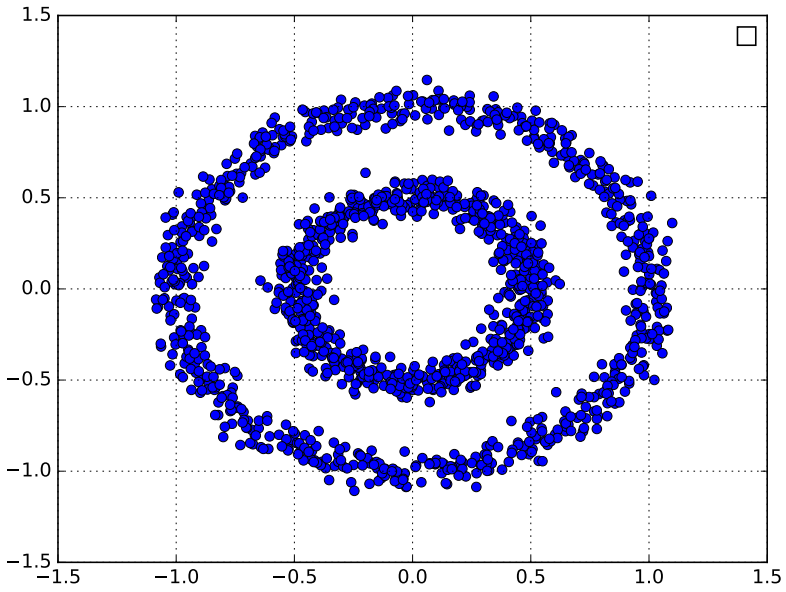


(b) Diversity based statistic $Q(k)$

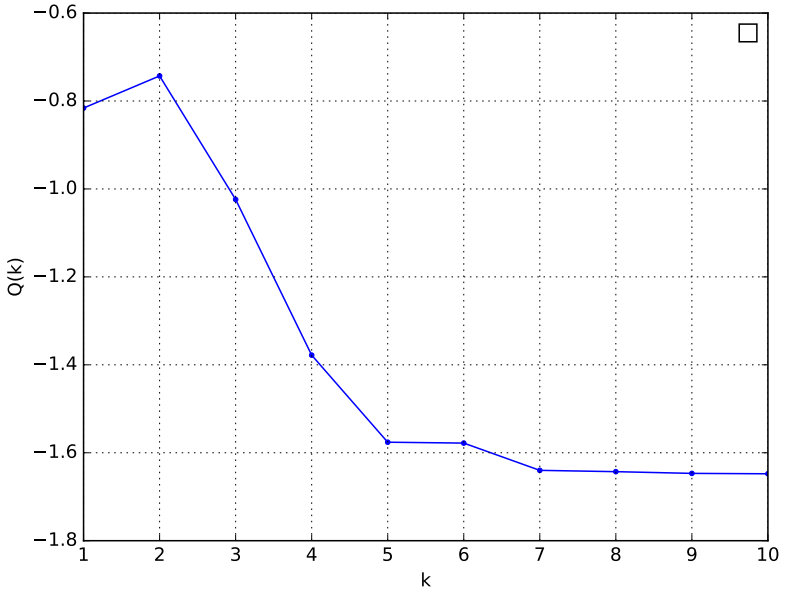
Figure 6.4: Experiments on the synthetic dataset with four clusters with different sizes and some random noise.

The items (data points) in each above cluster are all sampled from a particular standard multivariate normal distribution.

For each setting defined above, we generated 50 concrete datasets so as to carry out 50 simulation trials. Then we used the k-means clustering algorithm to divide the generated dataset into k clusters with k varying from 1 to 9. On the basis of the clustering results, we apply the diversity method and the



(a) Dataset

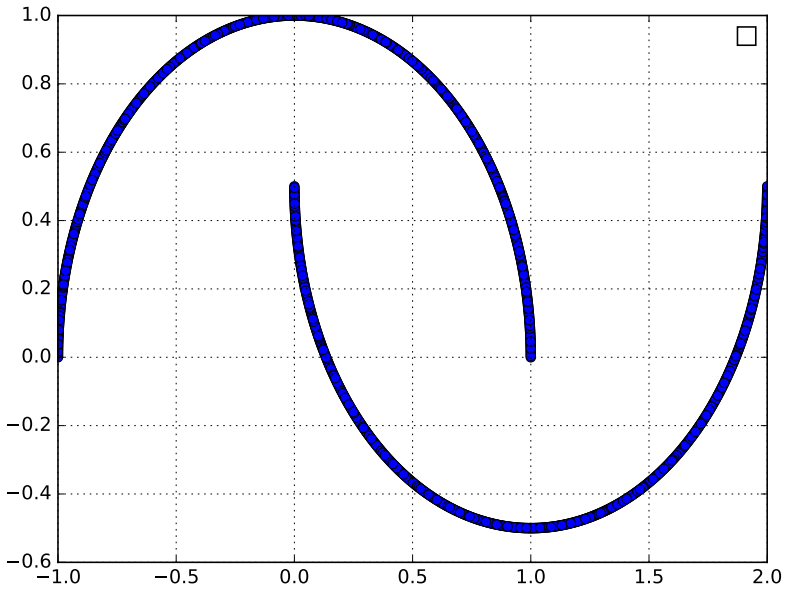


(b) Diversity based statistic $Q(k)$

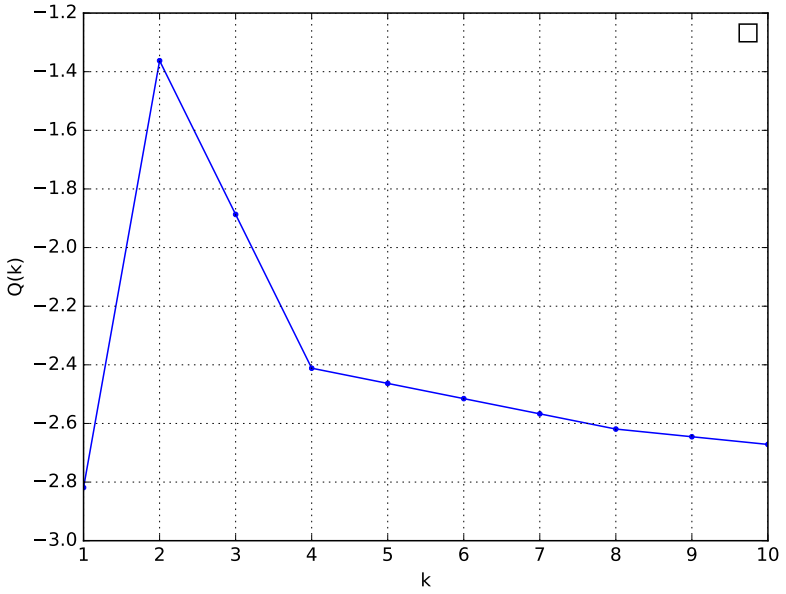
Figure 6.5: Experiments on the synthetic dataset with two ring-shape clusters.

other methods in comparison to make estimations about the actual number of clusters.

The results of the simulation study are summarised in Table 6.1. Each number in the table shows how many times a particular method detected the number of clusters mentioned in its column header. In the 1st case where there is little noise, all the methods perform almost equally well. In the 2nd case where there is a lot of noise, it can be clearly seen that the diversity methods



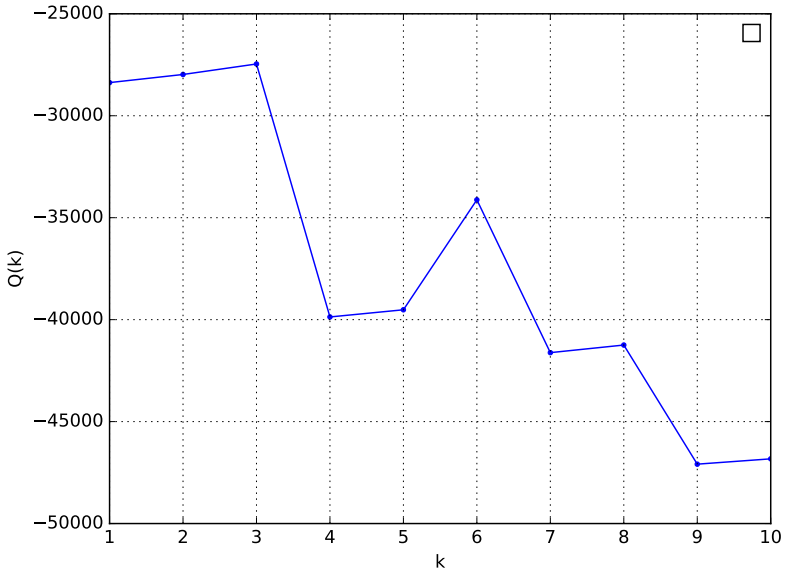
(a) Dataset



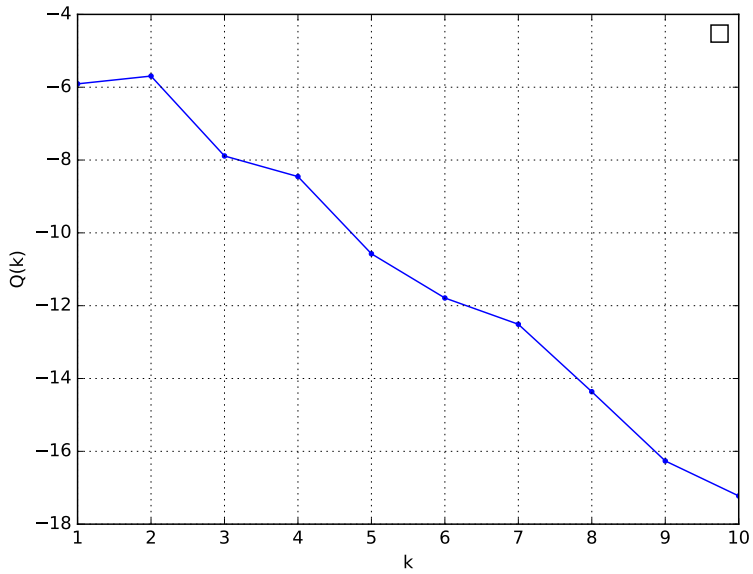
(b) Diversity based statistic $Q(k)$

Figure 6.6: Experiments on the synthetic dataset with two moon-shape clusters.

outperforms all the other methods significantly. In the 3rd and 4th case, the diversity method performs best with near-perfect accuracy, closely followed by the gap-statistic method (which is widely regarded as the state-of-the-art method).



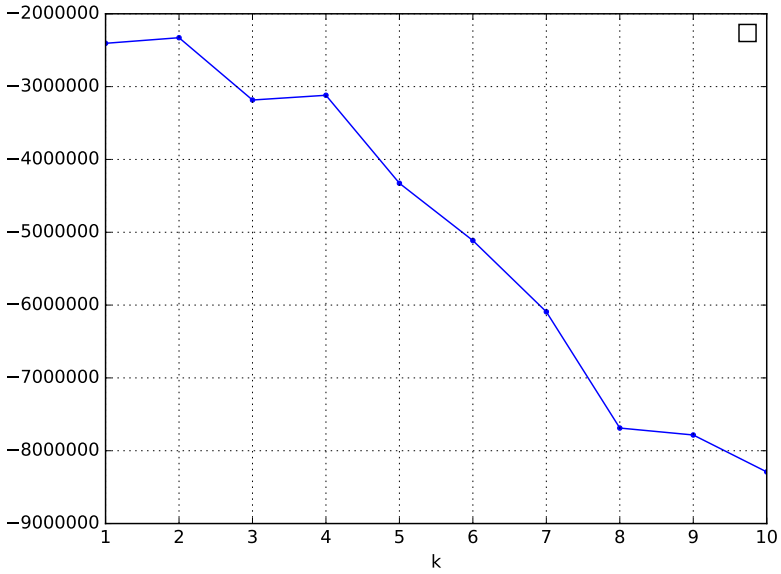
(a) Wine: $m = 13$, $k^* = 3$.



(b) Breast Cancer: $m = 9$, $k^* = 2$.

6.2 A META-EVALUATION OF EVALUATION METHODS FOR DIVERSIFIED SEARCH

For the evaluation of diversified search results, a number of different methods have been proposed in the literature. Prior to making use of such evaluation methods, it is important to have a good understanding of how diversity and relevance contribute to the performance metric of each method. In this paper, we use the statistical technique ANOVA to analyse and compare three representative evaluation methods for diversified search, namely α-nDCG, MAP-IA,



(c) Thyroid Disease: $m = 5$, $k^* = 2$.

Figure 6.6: Experimental results on three real-world datasets from UCI Machine Learning Repository, where m and k^* are the number of features/dimensions and the actual number of clusters respectively in the corresponding dataset.

and ERR-IA, on the TREC-2009 Web track dataset. It is shown that the performance scores provided by those evaluation methods can indeed reflect two crucial aspects of diversity — richness and evenness — as well as relevance, though to different degrees.

6.2.1 Overview

The same query could be submitted to a search engine by users from different backgrounds and with different information needs. When this occurs, the search engine should present users with relevant and diversified results that can cover multiple aspects or subtopics of the query. For more than a decade, there has been a surge of research in the diversification of search results [14, 96, 74, 45]. The main objective of such research is to deal with the ambiguity of query or the multiplicity of user intent.

To evaluate the performance of diversified search, a variety of metrics have been proposed in recent years, such as α-nDCG [21], MAP-IA [3], and ERR-IA [17] which generalise the corresponding traditional IR metrics [53] to capture both the *diversity* and the *relevance* of search results. Here, we aim to in-

Table 6.1: Experimental results on synthetic data showing how many times out of 50 simulation trials a particular method estimated the number of clusters to be $\widehat{k}$, where the column corresponding to the correct number of clusters is annotated with *.

Method	*Estimates of the following numbers of clusters* $\widehat{k}$								
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
(a) *Ground truth: 4 clusters (relatively clean)*									
elbow	0	0	1	49*	0	0	0	0	0
silhouette	0	0	0	50*	0	0	0	0	0
Caliński-Harabasz	0	0	0	50*	0	0	0	0	0
gap-statistic	0	0	0	50*	0	0	0	0	0
diversity	0	0	0	50*	0	0	0	0	0
(b) *Ground truth: 4 clusters (relatively noisy)*									
elbow	0	0	5	29*	16	0	0	0	0
Caliński-Harabasz	0	0	1	0*	49	0	0	0	0
silhouette	0	0	0	39*	11	0	0	0	0
gap-statistic	0	0	0	14*	36	0	0	0	0
diversity	0	0	0	48*	2	0	0	0	0
(c) *Ground truth: 5 clusters*									
elbow	0	1	0	5	44*	0	0	0	0
Caliński-Harabasz	0	7	0	6	37*	0	0	0	0
silhouette	0	2	0	9	39*	0	0	0	0
gap-statistic	0	0	0	0	48*	2	0	0	0
diversity	0	0	0	1	49*	0	0	0	0
(d) *Ground truth: 6 clusters*									
elbow	0	0	0	0	8	42*	0	0	0
Caliński-Harabasz	0	6	0	0	8	36*	0	0	0
silhouette	0	0	0	0	12	38*	0	0	0
gap-statistic	0	0	0	0	0	49*	1	0	0
diversity	0	0	0	0	0	50*	0	0	0

vestigate exactly how the above mentioned three representative performance metrics for diversified search are determined by diversity and relevance, using the Analysis of Variance (ANOVA) [29].

6.2.2 Related Work

The widely used IR performance metric nDCG [42] measures the accumulated usefulness ("gain") of the ranked result list with the gain of each relevant

document discounted at lower positions. Clarke et al. proposed its extended version α-nDCG [21] to evaluate diversified search results. It takes into account not only the position at which a relevant document is ranked but also the subtopics contained in that document, and uses a parameter $\alpha \in [0, 1)$ to control the severity of redundancy penalisation. Specifically, α-nDCG for the top-k search results is the discounted cumulative gain α-DCG[k] normalised by its "ideal" value, and DCG[k] can be calculated as:

$$\alpha\text{-DCG}[k] = \sum_{i=1}^{k} \frac{\sum_{s=1}^{N} g_{i,s}(1-\alpha)^{\sum_{j=1}^{i-1} g_{j,s}}}{\log_2(i+1)} \ , \tag{6.7}$$

where N is the total number of distinct subtopics, and $g_{i,s}$ is the human judgement for whether subtopic s is present or not in document i.

Agrawal et al. [3] proposed an approach to generalising traditional IR performance metrics for the search results of a query with multiple subtopics (user intents). The idea is to calculate the given performance metric for each subtopic separately, and then aggregate those scores based on the probability distribution of subtopics for the query. Extending the traditional IR performance metrics MAP [53] and ERR [18] in this way, we get their diversified versions:

$$\text{MAP-IA} = \sum_{s=1}^{N} P(s) \cdot \text{MAP}_s \quad \text{and} \quad \text{ERR-IA} = \sum_{s=1}^{N} P(s) \cdot \text{ERR}_s \ , \tag{6.8}$$

where N is the total number of distinct subtopics, $P(s)$ is the probability or weight of subtopic s, while MAP_s and ERR_s are the MAP and ERR scores for subtopic s respectively.

The previous studies most similar to our work are those from Clarke et al. [20] and Chandar et al. [16] which attempt to compare evaluation methods in the context of the diversified search. The former assumes simple cascade models of user behaviour, while the latter measures diversity just by the subtopic recall — *s-Recall* [93] — which may not reveal the full picture of diversity.

6.2.3 Meta-Evaluation

Factors for Meta-Evaluation of Evaluation Methods for Diversified Search

To examine the diversity of search results for a query, it is important to consider not only the number of distinct subtopics but also the relative abundance

of the subtopics present in the search result set. We use the concepts of *richness* and *evenness* as described in Section 2.2

Formally, we define the two measures, *richness* and *evenness*, in the context of the diversified search, as follows. The richness of the search result set for a query (topic) could be just defined as the number of distinct subtopics appeared in the set. In order to make the value of richness comparable across queries, we choose to use not the absolute number of distinct subtopics but the relative proportion of distinct subtopics:

$$richness = R/N \ , \tag{6.9}$$

where R is the number of distinct subtopics covered by the given search result set for a query, while N is the total number of distinct subtopics relevant to that query. This proportionate version of richness is actually equivalent to the *s-Recall* proposed by Zhai et al. [93]. The value of (proportionate) richness is obviously between 0 and 1. The evenness of the search result set for a query (topic) refers to how close in numbers each subtopic in the set is, i.e. it quantifies how evenly the search results are spread over the subtopics. To implement the *evenness* we use Eq. (2.3). For diversity index in evenness here, we use the well-known *Inverse Simpson's diversity index* as defined in Eq. (2.4).

For the purpose of assessing the *relevance* of search results, we can simply use the Precision@k measure [53], as in [16].

Data for Meta-Evaluation of Evaluation Methods for Diversified Search

The dataset used for our experiments comes from TREC-2009 Web track diversity task [19] which have also been used in previous studies [20, 16]. This dataset includes 50 topics, each of which consists of a set of subtopics representing different user needs.

Experiments for Meta-Evaluation of Evaluation Methods for Diversified Search

The evaluation methods for diversified search, including α-nDCG, MAP-IA, and ERR-IA, must be able to capture not only the relevance of search results but also the diversity of search results in terms of both richness and evenness.

The statistical technique, Analysis of Variance (ANOVA) [29], provides the perfect tool to gain insight into how each of these three factors (richness, evenness, and relevance) contributes to the overall performance measured by an evaluation method.

In our experiments, the dependent variable for the ANOVA would be the performance score given by α-nDCG[1], MAP-IA, or ERR-IA. Regarding the independent variables (richness, evenness, and relevance), since the real IR system outputs submitted to the TREC-2009 Web track could not account for all the possible scenarios that we would like to investigate, we generated a number of synthetic search result sets via a simulation process similar to the "*Rel+Div*" setting in [16]. Given a query (topic) in our dataset, we randomly sampled 10 documents form the full *qrels* file [19] to create such artificial document rankings that satisfy one of the $3^3 = 27$ different experimental conditions for top-10 search results: low/medium/high *richness*, low/medium/high *evenness*, and low/medium/high *relevance*, where the category labels low, medium, and high correspond to the value ranges 0.0–0.3, 0.3–0.6, and 0.6–1.0 respectively. The simulation process would continue until for each of the 50 queries (topics) we had generated 10 search result sets (rankings) per experimental condition. Therefore, the ANOVA for each evaluation method would have $50 \times 10 \times 27 = 13500$ data points to analyse.

Results of Meta-Evaluation of Evaluation Methods for Diversified Search

The statistical significance results of the ANOVA are shown in Table 6.2. It can be seen that all those performance metrics, α-nDCG, MAP-IA, and ERR-IA, would be influenced heavily by the individual factors — richness, evenness, and relevance — with almost zero p-values, but not so much by their interactions. This confirms that the chosen three factors are relatively independent (untangled) aspects of a system's performance for diversified search.

Furthermore, Table 6.3 shows the variance decomposition results of the ANOVA, where SSE stands for the sum of squared errors. It seems that MAP-IA reflects more richness than the other two performance metrics, as the change of richness accounts for 13% of the total variability in MAP-IA

[1]The parameter α for α-nDCG was set to 0.5, the default value used in the TREC-2009 Web track diversity task.

Table 6.2: The statistical significance results of the ANOVA.

Component	α-nDCG		MAP-IA		ERR-IA	
	F	p-value	F	p-value	F	p-value
richness	362.4	0.00	590.9	0.00	253.7	0.00
evenness	480.0	0.00	521.7	0.00	282.7	0.00
relevance	465.7	0.00	285.0	0.00	397.0	0.00
richness * *evenness*	10.8	0.00	2.9	0.03	0.9	0.46
richness * *relevance*	3.5	0.01	5.3	0.00	5.8	0.00
evenness * *relevance*	4.3	0.00	0.3	0.91	3.2	0.01
richness * *evenness* * *relevance*	0.8	0.53	1.4	0.24	2.4	0.05

Table 6.3: The variance decomposition results of the ANOVA.

Component	α-nDCG		MAP-IA		ERR-IA	
	SSE	(%)	SSE	(%)	SSE	(%)
richness	13.1	(8%)	8.2	(13%)	2.0	(6%)
evenness	17.4	(11%)	7.2	(11%)	2.3	(7%)
relevance	16.9	(10%)	3.9	(6%)	3.2	(9%)
richness * *evenness*	0.6	(0%)	0.1	(0%)	0.0	(0%)
richness * *relevance*	0.3	(0%)	0.1	(0%)	0.1	(0%)
evenness * *relevance*	0.3	(0%)	0.0	(0%)	0.1	(0%)
richness * *evenness* * *relevance*	0.1	(0%)	0.0	(0%)	0.0	(0%)
residual	117.0	(71%)	44.6	(70%)	25.9	(77%)

which is substantially higher than 8% in α-nDCG and 6% in ERR-IA. On the other hand, evenness is probably reflected better by α-nDCG or MAP-IA than ERR-IA, as the change of evenness accounts for 11% of the total variability in α-nDCG and MAP-IA but only 7% in ERR-IA. In terms of relevance, α-nDCG looks the most accurate indicator, because 10% of its total variability is attributed to the change of relevance, which is followed by 9% in ERR-IA and 6% in MAP-IA. The "residual" component which comprises everything about the performance metric unexplained by the proposed independent variables (factors) occupies a high proportion of the total variability, which suggests that the difficulty of the query (topic) and also the specific ranking algorithm still play the major roles in determining performance scores.

6.3 CONCLUSION

In this chapter, we used a novel *diversity* based approach to the problem of estimating the number of clusters in a dataset. To our knowledge, the underlying connection between diversity and clustering has not been revealed before in research literature.

Specifically, we show that the difference between the global diversity of clusters and the sum of each cluster's local diversity of their members can be used as an effective indicator of the optimality of the number of clusters, where the diversity is measured by Rao's quadratic entropy. A notable advantage of our proposed method is that it encourages *balanced* clustering by taking into account both the sizes of clusters and the distances between clusters. In other words, it is less prone to very small "outlier" clusters than existing methods.

Our extensive experiments on both synthetic and real-world datasets (with known ground-truth clustering) have demonstrated that our proposed method is robust to clusters of different sizes and shapes, and it is more accurate than existing methods (including elbow, Caliński-Harabasz, silhouette, and gap-statistic) in terms of finding out the true number of clusters.

Afterwards, we used richness, evenness and precision of search results to meta-evaluate the evaluation of diversified search results. Our experiments using ANOVA have indicated that richness, evenness, and relevance could be well differentiated by three representative evaluation methods for diversified search — α-nDCG, MAP-IA and ERR-IA — We further observed that each of these evaluation methods has a different level of contribution for the proportion of total variability, in the components of richness, evenness, and relevance.

CHAPTER 7

CONCLUSIONS

In this chapter, we summarise and present research contributions in the thesis. In Section 7.1, contributions of this thesis are presented. The thesis concludes in Section 7.2, with a discussion of the prospects for future research and development of diversity in the Web.

7.1 CONTRIBUTIONS

We found the current issues in the Web which required to have a mechanism to measure and compare diversity over the Web. In Chapter 1 we mentioned that there arises an ambiguity when the information required by users is limited to a few key-words which does not reflect the actual needs of these users. This ambiguity in the need for actual information can be solved with the help of presenting the diversified results set. The analysis of diversity has been extensively studied and its concepts are very well established in the field of ecology, where the types of interest are species in a certain region. We introduced these methods namely, Inverse Simpson's index, Shannon's diversity index and Rao's quadratic diversity, to analyse the diversity in the Web.

In Chapter 2, we discussed the research literature about introducing diversity in the Web. We discussed the extrinsic and intrinsic approaches to deal with the diversity when the queries posted by users are either ambiguous or underspecified. Afterwards, we mentioned well-known methods to reduce redundancy and introduce novelty in the result set for any query. We further went into review and compare the methods, e.g. MMR, ERR and α–nDCG

are a few among others, which focused on introducing diversity in results with respect to the multiplicity of topics for a query.

In Chapter 3, we introduced a novel method to analyse the diversity of Web search results in two well-known Web search engines namely, Google and Bing. After reviewing the current research literature, to the best of our knowledge this method to analyse the diversity in the Web is not explored before. Firstly, we theoretically observed that the diversity in Web search results is determined by the Zipfian distribution of websites. Secondly, we compared Google and Bing in terms of differences in richness, evenness and diversity, with two methods Inverse Simpson's index and Shannon's diversity index, of organic and advert websites. Along with that we also compared the topic diversity in both aforementioned Web search engines. We showed that the differences in diversity in both Web search engines are statistically significant by using the non-parametric statistical significance test.

In Chapter 4 we showed how to predict the lifetime of popularity of a query, i.e. how long a query remains in Google top charts. For that we investigated the diversity of queries coverage in Google and Bing. The results of the diversity of queries coverage, i.e. diversity and evenness, are used as covariates in Cox proportional hazard regression model for predicting the lifetime of popularity of a query. We observed the highest value of concordance index, which represents how well our model predicts the lifetime of a query, when we used diversity related factors, i.e. diversity and evenness, along with other factors mentioned in Table 4.5.

It is a requirement of Inverse Simpson's index and Shannon's diversity index that all the items should be in separate categories, i.e. the categories should not be overlapping. In Chapter 5 we showed how to implement the diversity measures when the categories are overlapping. We introduced two separate methods to deal with overlapping categories. Afterwards, we dealt with diversity with respect to time. We showed how we can get some meaningful information, e.g, trend and seasonality, by using time series decomposed components, which is not instantly obvious from the diversity values of historical data.

Last but not the least in Chapter 6 we further show that how to utilise and implement the diversity measures in two applications over the Web. we

show that how we can estimate the number of balanced clusters using a novel diversity based approach. We used Rao's quadratic entropy, which uses size and distance between clusters, to estimate the number of clusters in a dataset. Afterwards, we meta-evaluate the evaluation of diversified search results. We show how the well known evaluation methods, e.g. α–nDCG, MAP-IA and ERR-IA, can differentiate between different levels of richness, evenness and relevance of diversified Web search results.

7.2 FUTURE WORK

Although we have investigated only into the analysis of diversity in the Web. It would be interesting to see the relationship of diversity with overall users' satisfaction. For example, we can investigate into the relationship of relevance with diversity and analyse to see what level or value of diversity is better suited for a Web search engine, so that it satisfies all the users with different backgrounds and with different needs.

To investigate if a few well-established websites dominate the Web search engines, we analysed the Web search engines for the diversity of sources. Other than source diversity in Web search results, there is another interesting aspect of diversity, i.e. the type of source diversity. For example, we can classify the documents in Web search results into type of source, e.g. News, Blog, Social network, Videos, Images and Recreational, and apply the same diversity models, e.g. Inverse Simpson's index and Shannon's diversity index, to better understand the diversity in another dimension, i.e diversity of the type of source in Web search results.

We have shown how to analyse the diversity in different datasets, and in different scenarios, over the Web. In particular, we analysed the diversity in Web search engine results and its queries, dataset for movies and the role of diversity in clustering textual datasets. In our experiments we compared and analysed two English language Web search engines, Google and Bing. It would also be interesting and more informative to include other language Web search engines or cover multiple languages and see how it effects the diversity. An interesting aspect of diversity can be shown by looking into what is the role of diversity in the various other types of datasets and user-oriented applications over the Web, such as recommender systems and social media networks.

For example, it would greatly improve the user satisfaction, and ultimately the whole business model, if we could know the relationship between diversity in recommended products and user buying pattern over an e-commerce website, such as amazom.com. On the other hand, there is a long debated issue of bias in the conventional media sources, such as newspapers and television. This issue can be better addressed through investigating and comparing the topic diversity for news items in conventional news media sources, e.g. NewYorkTimes, TheGuardian or DailyMail, with the news items in social media networks, such as Twitter or reddit, and analyse for any bias among these news sources.

BIBLIOGRAPHY

[1] Search engine market share. `https://netmarketshare.com/search-engine-market-share.aspx`. Accessed: 01-June-2018.

[2] L. A. Adamic. Zipf, Power-Laws, and Pareto — A Ranking Tutorial. `http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html`, 2000.

[3] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *Proceedings of the 2nd International Conference on Web Search and Data Mining*, WSDM, pages 5–14, Barcelona, Spain, 2009.

[4] M. Andersen and H. Taylor. *Sociology: The Essentials*. Cengage Learning, 2012.

[5] C. Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion Books, 2006.

[6] X. Arzoz. *Respecting Linguistic Diversity in the European Union*, volume 2. John Benjamins Publishing, 2008.

[7] K. Bache, D. Newman, and P. Smyth. Text-Based Measures of Document Diversity. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pages 23–31, Chicago, Illinois, USA, 2013.

[8] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query Recommendation Using Query Logs in Search Engines. In *International Conference on Extending Database Technology*, EDBT, pages 588–596, Heraklion, Crete, Greece, 2004.

[9] A. L. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.

[10] M. Begon, J. L. Harper, and C. R. Townsend. *Ecology: Individuals, Populations, and Communities, .* John Wiley & Sons, 3rd edition, 1996.

[11] P. Borlund. The Concept of Relevance in IR. *Journal of the American Society for information Science and Technology*, 54(10):913–925, 2003.

[12] E. Brynjolfsson, Y. J. Hu, and M. D. Smith. The Longer Tail: The Changing Shape of Amazon's Sales Distribution Curve. *SSRN Electronic Journal*, 2010.

[13] T. Caliński and J. Harabasz. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[14] J. G. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 335–336, Melbourne, Australia, 1998.

[15] P. Castells, J. Wang, R. Lara, and D. Zhang. Introduction to the Special Issue on Diversity and Discovery in Recommender Systems. *ACM Transactions on Intelligent Systems and Technology*, 5(4):52, 2014.

[16] P. Chandar and B. Carterette. Analysis of Various Evaluation Measures for Diversity. In *Proceedings of the Diversity in Document Retrieval Workshop*, DDR, pages 21–28, Dublin, Ireland, 2011.

[17] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-Based Diversification of Web Search Results: Metrics and Algorithms. *Information Retrieval Journal*, 14(6):572–592, 2011.

[18] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM, pages 621–630, Hong Kong, China, 2009.

[19] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 Web Track. In *Proceedings of the 18th Text REtrieval Conference*, TREC, pages 17–20, Gaithersburg, MD, USA, 2009.

[20] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proceedings of the 4th International Conference on Web Search and Data Mining*, WSDM, pages 75–84, Hong Kong, China, 2011.

[21] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 659–666, Singapore, Singapore, 2008.

[22] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, 2009.

[23] G. B. Coleman and H. C. Andrews. Image Segmentation by Clustering. *Proceedings of the IEEE*, 67(5):773–785, 1979.

[24] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[25] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An Experimental Comparison of Click Position-Bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM, pages 87–94, New York, NY, USA, 2008.

[26] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*, volume 283. Addison-Wesley Reading, 2010.

[27] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query Expansion by Mining User Logs. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):829–839, 2003.

[28] K. Denecke. *Web Search Engine Research*, chapter 6 Diversity-Aware Search: New Possibilities and Challenges for Web Search, pages 139–162. Emerald Group Publishing Limited, 2012.

[29] G. Gamst, L. S. Meyers, and A. J. Guarino. *Analysis of Variance Designs: A Conceptual and Computational Approach With SPSS and SAS*. Cambridge University Press, 2008.

[30] F. Giunchiglia, V. Maltese, D. Madalli, A. Baldry, C. Wallner, P. Lewis, K. Denecke, D. Skoutas, and I. Marenzi. Foundations for the Representation of Diversity, Evolution, Opinion and Bias. Technical report, University of Trento, 2009.

[31] A. D. Gordon. Null Models in Cluster Validation. In P. D. Gaul W., editor, *From Data to Knowledge. Studies in Classification, Data Analysis, and Knowledge Organization*, pages 32–44. Springer, Berlin, Heidelberg, 1996.

[32] C. Goutte, P. Toft, E. Rostrup, F. Å. Nielsen, and L. K. Hansen. On Clustering fMRI Time Series. *NeuroImage*, 9(3):298–310, 1999.

[33] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, 15(4):361–387, 1996.

[34] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[35] O. C. Herfindahl. *Concentration in the US Steel Industry*. PhD thesis, Columbia University, 1950.

[36] M. O. Hill. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, 54(2):427–432, 1973.

[37] A. O. Hirschman. *National Power and the Structure of Foreign Trade*. University of California Press, Berkeley and Los Angeles, 1945.

[38] S. H. Hurlbert. The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology*, 52(4):577–586, 1971.

[39] A. K. Jain. Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[40] B. J. Jansen, D. L. Booth, and A. Spink. Determining the User Intent of Web Search Engine Queries. In *Proceedings of the 16th International Conference on World Wide Web*, WWW, pages 1149–1150, banff alberta canada, 2007.

[41] B. J. Jansen, A. Spink, and T. Saracevic. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing & Management*, 36(2):207–227, 2000.

[42] K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems, TOIS*, 20(4):422–446, 2002.

[43] L. Jost. Entropy and Diversity. *Oikos*, 113(2):363–375, 2006.

[44] D. J. Ketchen Jr and C. L. Shook. The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, 17(6):441–458, 1996.

[45] S. K. Kingrani, M. Levene, and D. Zhang. Diversity Analysis of Web Search Results. In *Proceedings of the ACM Web Science Conference*, WebSci, pages 43:1–43:2, Oxford, UK, 2015.

[46] D. E. Knuth. *The Art of Computer Programming.* Addison-Wesley, 3rd edition, 1997.

[47] A. N. Langville and C. D. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings.* Princeton University Press, 2011.

[48] J. G. Lee, S. Moon, and K. Salamatian. Modeling and Predicting the Popularity of Online Contents With Cox Proportional Hazard Regression Model. *Neurocomputing*, 76(1):134–145, 2012.

[49] M. Levene. *An Introduction to Search Engines and Web Navigation.* Wiley-Blackwell, 2nd edition, 2010.

[50] M. Lichman. UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml`, 2013.

[51] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[52] A. E. Magurran. *Ecological Diversity and Its Measurement.* Princeton University Press, 1988.

[53] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[54] G. W. Milligan and M. C. Cooper. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2):159–179, 1985.

[55] E. Mills. Aol sued over web search data release. CNET news. `https://www.cnet.com/news/aol-sued-over-web-search-data-release`, 2006. Accessed: 01-June-2018.

[56] B. Mirkin. *Clustering: A Data Recovery Approach*. CRC Press, 2012.

[57] N. J. D. Nagelkerke. A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78(3):691–692, 1991.

[58] M. E. J. Newman. Power Laws, Pareto Distributions and Zipf's Law. *Contemporary Physics*, 46(5):323–351, 2005.

[59] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, NIPS, pages 849–856, Vancouver, Canada, 2002.

[60] S. E. Page. *Diversity and Complexity*. Princeton University Press, 2010.

[61] E. Pariser. *The Filter Bubble: What the Internet is Hiding From You*. Penguin UK, 2011.

[62] E. C. Pielou. *An Introduction to Mathematical Ecology*. Wiley-Interscience, 1969.

[63] J. Pitkow et al. Personalized search: A content computer approach may prove a breakthrough in personalized search efficiency. *Communications of the ACM*, 45(9):50–55, 2002.

[64] D. M. W. Powers. Applications and Explanations of Zipf's Law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP/CoNLL, pages 151–160, Sydney, Australia, 1998.

[65] K. Purcell, J. Brenner, and L. Rainie. Search Engine Use 2012. `http://pewinternet.org/Reports/2012/Search-Engine-Use-2012.aspx`, 2012.

[66] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, Diversity and Interdependent Document Relevance. In *SIGIR Forum*, volume 43, pages 46–52, 2009.

[67] F. Radlinski and S. Dumais. Improving Personalized Web Search Using Result Diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 691–692, Seattle, Washington, USA, 2006.

[68] F. Radlinski, R. Kleinberg, and T. Joachims. Learning Diverse Rankings With Multi-Armed Bandits. In *Proceedings of the 25th International Conference on Machine Learning*, ICML, pages 784–791, Helsinki, Finland, 2008.

[69] C. R. Rao. Diversity and Dissimilarity Coefficients: A Unified Approach. *Theoretical Population Biology*, 21(1):24 – 43, 1982.

[70] J. J. Rocchio. Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323, 1971.

[71] M. L. Rosenzweig. *Species Diversity in Space and Time*. Cambridge University Press, 1995.

[72] P. J. Rousseeuw and L. Kaufman. *Finding Groups in Data*. Wiley Online Library, 1990.

[73] S. Salvador and P. Chan. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In *16th IEEE International Conference on Tools With Artificial Intelligence*, ICTAI, pages 576–584, Boca Raton, Florida, USA, 2004.

[74] R. L. T. Santos, C. Macdonald, and I. Ounis. Search Result Diversification. *Foundations and Trends in Information Retrieval*, 9(1):1–90, 2015.

[75] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting Query Reformulations for Web Search Result Diversification. In *Proceedings of the 19th International Conference on World Wide Web*, WWW, pages 881–890, 2010.

[76] C. Sha, K. Wang, D. Zhang, X. Wang, and A. Zhou. Optimizing Top-K Retrieval: Submodularity Analysis and Search Strategies. In *Proceeedings of the 15th International Conference on Web-Age Information Management*, WAIM, pages 18–29, Macau, China, 2014.

[77] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423 & 623–656, 1948.

[78] E. H. Simpson. Measurement of Diversity. *Nature*, 163(4148), 1949.

[79] D. Skoutas, E. Minack, and W. Nejdl. Increasing Diversity in Web Search Results. In *Proceedings of Web Science 2010*, WebSci, Raleigh, NC, USA, 2010.

[80] A. R. Solow. A Simple Test for Change in Community Structure. *Journal of Animal Ecology*, 62(1):191–193, 1993.

[81] R. Song, Z. Luo, J.-R. Wen, Y. Yu, and H.-W. Hon. Identifying Ambiguous Queries in Web Search. In *Proceedings of the 16th International Conference on World Wide Web*, WWW, pages 1169–1170, Banff, Alberta, Canada, 2007.

[82] A. Stirling. A General Framework for Analysing Diversity in Science, Technology and Society. *Journal of the Royal Society Interface*, 4(15):707–719, 2007.

[83] R. L. Thorndike. Who Belongs in the Family? *Psychometrika*, 18(4):267–276, 1953.

[84] R. Tibshirani, G. Walther, and T. Hastie. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[85] G. Van Rossum and F. L. Drake Jr. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[86] S. Vigna. Fibonacci Binning. *Computing Research Repository, CoRR*, abs/1312.3749, 2013.

[87] Y. Virkar and A. Clauset. Power-Law Distributions in Binned Empirical Data. *The Annals of Applied Statistics*, 8(1):89–119, 2014.

[88] E. M. Voorhees. The TREC-8 Question Answering Track Report. In *Trec*, volume 99, pages 77–82, 1999.

[89] E. M. Voorhees and H. T. Dang. Overview of the TREC 2005 Question Answering Track. In *Proceedings of the Fourteenth Text REtrieval Conference*, TREC, pages 15–18, Gaithersburg, Maryland, USA, 2005.

[90] J. Wang and J. Zhu. Portfolio Theory of Information Retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 115–122, Boston, MA, USA, 2009.

[91] M. Wedel and W. A. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. Springer, 2012.

[92] L. Xu and M. I. Jordan. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, 8(1):129–151, 1996.

[93] C. Zhai, W. W. Cohen, and J. D. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pages 10–17, Toronto, Canada, 2003.

[94] W. Zhou and W. Duan. Online User Reviews, Product Variety, and the Long Tail: An Empirical Investigation on Online Software Downloads. *Electronic Commerce Research and Applications*, 11(3):275–289, 2012.

[95] G. Zuccon and L. Azzopardi. Using the Quantum Probability Ranking Principle to Rank Interdependent Documents. In *Proceedings of the 32nd European Conference on IR Research*, ECIR, pages 357–369, Milton Keynes, UK, 2010.

[96] G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-K Retrieval Using Facility Location Analysis. In *Proceedings of the 34th European Conference on IR Research*, ECIR, pages 305–316, Barcelona, Spain, 2012.

APPENDIX A

LIST OF PUBLICATIONS

The papers resulting from the author's PhD research are as follows.

- Suneel Kumar Kingrani, Mark Levene, and Dell Zhang. Diversity Analysis of Web Search Results. In *Proceedings of the ACM Web Science Conference (WebSci)*, pp. 43:1–43:2, Oxford, UK, 28 Jun – 1 Jul 2015.
- Suneel Kumar Kingrani, Mark Levene, and Dell Zhang. A Meta-Evaluation of Evaluation Methods for Diversified Search. In *Proceedings of the 40th European Conference on Information Retrieval (ECIR)*, pp. 550–555, Grenoble, France, Mar 2018.
- Suneel Kumar Kingrani, Mark Levene, and Dell Zhang. Estimating the Number of Clusters Using Diversity. *Artificial Intelligence Research*, 7(1), pp. 15–22, 2018.
- Suneel Kumar Kingrani, Mark Levene, Dell Zhang. Analysing the Diversity of Web Search Results. Submitted to *World Wide Web (WWW)*, Springer.

APPENDIX B

KEY TERMS USED IN THESIS

- Adverts Websites

The paid advertisement websites which appear alongside the main results of any search query in a Web search engine.

- ANOVA

Analysis of variance is a statistical method in which we test groups and see if there are differences between them.

- C-index

It is the probability of concordance between predicted and observed responses.

- Clustering

It is an statistical technique in which similar objects are grouped (clustered) together.

- Cox Proportional Hazard Regression Model

It is an statistical model which is being used to explore the relationship between the "survival" (in presence of censored data) of a subject and the explanatory variables. Where the response variable is the hazard function at a given time t.

- Evenness

Distribution of species or how evenly the individuals are distributed among the species in a jungle.

- Fibonacci Binning

It is a simple logarithmic binning technique in which bins are sized like the Fibonacci numbers.

- Inverse Simpson's Diversity

It is the inverse of the probability that two individuals chosen at random belong to same species.

- Organic Websites

The websites which appear as the main result of any search query over a Web search engine.

- Power Law

It is the relationship between two quantities in which one quantity, x varies as the power of the other quantity, y.

- Power Law with Exponential Cut-off

In this Power Law holds for small values of x, but then turns smoothly into a declining exponential function for large values of x. The exponential, large-x tail drops faster than the Power Law.

- Queries and Related Queries

The queries are posted by user to find to find the required information over the Web. Related queries are provided by Web search engines, usually at the bottom of the Web search results.

- Rao's Quadratic Diversity

It is the expected distance between two randomly chosen individuals from all the species in a jungle.

- Richness

Number of distinct species, N, in a jungle.

- Shannon's Diversity

It is the uncertainty in predicting the species type of an individual who is chosen at random.

- Zipf's Law

It is one of a family of related Power Law. In which one quantity, y varies as the power of its rank.