The initial explosive growth of the web, now often referred to as Web 1.0 [145], led to a huge increase in information available online: companies created their own web presence, newspapers began offering news articles to readers online, governments started to inform their citizens using websites, and many more organizations allowed online users to find at least the most basic information online. Two main characteristics of this initial information boom are (i) the content creators (webmasters or online editors) were specialized positions within organizations and (ii) the involvement of web users was mainly restricted to consuming information.

Starting in the twenty first century, the web experienced another phase of explosive growth and this time web users were the ones to cause this growth. A large number of platforms became available for users to publish information, communicate with others, connect to like-minded, and share anything that they wanted to share. Today, we still have not reached the point of saturation: new platforms are being introduced all the time, and some of these manage to attract huge numbers of users in a relatively short amount of time. To give an idea of the types of platform that are available to users nowadays to share, connect, inform, and communicate, we list a few examples.

**Picture and video sharing:** Visual content created by individuals or companies can be made public; viewers of the content can comment on the items, but also add tags, even at a detailed level (one face in the picture, a few seconds in the video).

**Music:** Compose playlists to share with friends, tag bands and songs, see what others listen to, and construct and share your own music profile. Of course, music related social media also allow you to share your own music with the world.

**Mailing list:** Discussions are started by replying to earlier email messages. Postings on the list are usually stored online, creating an email archive. Mailing lists tend to be restricted to one topic (or domain), like soccer, digital cameras, or Moroccan culture.

**Forum:** Users can create a profile, start discussions, and contribute to these. Like mailing lists, forms are often devoted to one topic.

**Blog:** Often referred to as an online journal. Blogs allow users to easily share an experience or view, often with facilities to allow readers to comment on the initial
message, thus allowing some interaction between blogger (blog creator) and readers.

**Community question answering:** Allows users to ask a question, that fellow users can answer. Users can rate answers that are given, as a way of identifying the “best” answer.

**Collaborative knowledge source:** Facilitates the sharing of expertise. People can contribute to topics they know about, and together create an entry on a topic. Shared knowledge sources can be very extensive, with wide coverage.

**Social networking platform:** Has a range of possible uses (connect with friends, professionals, people with similar interests, etc.), but all evolve around the idea of discovering new people and making it easy to keep in touch. Often incorporates other platforms of sharing information.

**Microblog:** Allows users to give (close to) real-time updates of activities or thoughts. Messages are very short (∼140 characters) and typically aimed at a set of “followers.”

All of the platforms listed above are examples of **social media:** “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user generated content” [88]. Social media is a form of many-to-many communication. In principle, everyone can create content, which in turn can, in principle, be read by everyone else. However, to make the content available to everyone, people need to be able to identify the “right” pieces of content, or the “appropriate” content creators. That is, we need ways to intelligently access information in social media.

In the remainder of this thesis we focus on textual social media, and ignore other media types, like audio, video, and images. Although these media types are very valuable, and interesting from a research point of view, they are outside the scope of the research as we restrict ourselves to textual sources.

### 1.1 Information in Social Media

Why should we care about the information contained in social media? The short answer is: because such information gives rise to unique new types of information needs. To illustrate this, we give seven examples of information needs in social media.

**Marketing and sales:** Before buying a product, consumers often look for reviews of these products online, where large numbers of people share their experiences in the form of reviews, mainly in blogs and forums [24, 141]. For producers, identifying the most influential people who review their products may be very important: targeting this specific group, and trying to get them to “promote” their product may lead to an increase in uptake.

**Viewpoint research:** To get a better understanding of complex issues, it is often useful to look at the issue from different points of view. The large number of people
writing about what they think about a certain issue, makes it possible to collect these different viewpoints. Social media are a valuable source for collecting these viewpoints [68, 123]. The task is relevant for political analysts and journalists.

**Helpful answers:** You can hardly think of a problem that no-one else has had before, and solutions for these problems are available online. Various social media platforms, most notably mailing lists, forums, and community-QA sites, focus around problems and their solutions [23, 205, 209]. Offering access to the correct information that can lead to solving, or at least improving someone’s understanding of, a particular problem, proves to be very valuable, in a range of domains (medical, career choice, DIY, etc).

**Market research and product development:** Boosting sales is one thing, but researching the market to look for opportunities is another challenge that can make good use of access to social media [76, 83, 100, 148]. What features would people like to have in a product? How do they experience certain activities? What is the response to a new policy? Summarizing social media with regard to these questions leads to a very extensive type of market research.

**Intelligence and profiling:** With many people expressing opinions and views, and most of it relatively easy collectable, social media offer a wealth of information for intelligence [4, 47]. Intelligence agencies are particularly interested in gaining access to this information, to detect people who display, in some way, “interesting” behavior. Related to this is profiling of people [14]: Using social media to construct profiles of people. What are they interested in? What are their areas of expertise? Who are their friends? Who do they disagree with?

**News impact:** Not all news has an equal impact on people; some news stories are mostly ignored, whereas other stories generate a large volume of discussion. By looking at comments made on news articles or examining news related (micro) blog posts, news agencies can determine which stories appear to have more impact than others [173, 186, 188, 189]. This can be used in ad pricing or news paper lay-out decisions or simply to help understand people’s behavior regarding news.

**Influentials and experts:** When reading people’s messages, we might be more interested in an expert view on the topic, or we could recommend an expert to our friends as someone worth reading or listening to. Identifying experts in social media is an area worth exploring [30, 113, 215], just like the influentials: people who influence large groups of people. Being able to identify these allows companies to target specific users, and thereby reach a large audience [59, 94, 206].

Looking at these examples of information needs in social media, we observe that they revolve not just around relevance: we are not just concerned with finding the information objects that are about a given topic. Other criteria play an important role in determining which information is interesting to the information seeker: People need to be authorities or possess some level of expertise on a topic; information needs to be credible, it is not supposed to be a repetition of previously seen documents (novelty is important), and in fast-changing platforms, recency is an important aspect; finally, documents should
1. Introduction

contain opinions on a topic, or describe an experience. Many more ranking criteria exist and each of the ranking criteria is valid in its own right and possibly challenging. Still, we are almost always interested in these criteria after we have established that a document is about the topic of interest.

Social media are characterized by the lack of top-down rules and editors. Formal texts, like news articles and company messages, are usually checked by editors (e.g., to correct grammar and spelling errors) and written taking into account a set of top-down rules (e.g., how to refer to entities, maximum sentence length, clear writing style). These rules and editors make sure that formal texts have a certain quality level and are relatively easy to comprehend. Since social media platforms allow anyone to write whatever they feel like, in whatever form they want, we cannot give any assurance as to the quality of these messages. Social media texts are noisy: they contain spelling mistakes, grammatical errors, and creative language usage. The noisy character of the data in social media poses a large challenge to the information retrieval field.

The main motivation for the research in this thesis follows from the two preceding paragraphs: We want to enable intelligent access to, and analysis of, information contained in the noisy texts of social media. To this end, we need to determine topical relevance of social media documents, while countering the specific challenges posed by the noisy character of these documents.

1.2 Research Outline and Questions

We can visualize social media usage as done in Figure 1.1. The figure shows the usage of social media: a user, influenced by his environment, expresses himself on one of the platforms to which he is subscribed (e.g., microblog platform Twitter, \(^1\) social network Facebook, \(^2\) or blogservice Blogger\(^3\)). This leads to large numbers of messages on these platforms, all belonging to the same user. In the following paragraphs we briefly discuss the elements of this figure and proceed to our research questions.

The most important element of social media usage, as depicted in Figure 1.1, are people. As we can see, we can approach the user from two ways: (i) left to right, and (ii) right to left. In the case of (i), we search for people and characterize them by their presence on social media platforms. Approach (ii) starts with the texts published by the user and uses these to represent a person. By exploring what a user wrote, we can identify people with a certain level of interest in a given topic.

The second element we discuss are the messages created on social media platforms. Since messages in social media are often only short blurbs of text, not necessarily meant to convey a report on objective facts or events, we rather refer to them as utterances. Examples of utterances are blog posts, status updates, tweets, emails, questions, and forum messages. As mentioned before, utterances are characterized by their noisiness, a result from the lack of rules and editors in social media, something we need to take into account in our research. Another characteristic of utterances is the fact that they are embedded within a broader context, within the platform they belong to. What do

\(^1\)http://www.twitter.com  
\(^2\)http://www.facebook.com  
\(^3\)http://www.blogger.com
1.2. Research Outline and Questions

Figure 1.1: A user, influenced by what goes on around him, expresses himself on various social media platforms, resulting in “heaps” of social media utterances.

we mean by this? Imagine an utterance on a forum, i.e., a forum post; not only do we know the content of this post, we also know that it belongs to a discussion (or thread) regarding a topic. This discussion is part of a (sub-)forum, which in itself could be part of a community (e.g., a website, manuals, documents, community members). All these levels of context influence to some extent the content of the utterance.

Finally, we observe that social media platforms do not exist by themselves, but are surrounded by a real-world environment. Users of the platforms take note of this environment from, for example, news papers, television shows, social interaction, and other social media platforms. Being influenced by this environment, people may refer to this environment in their utterances. We observe, for example, that much of what people blog about is influenced by what happens in the news [138] and that “popular” people (highly frequent queries) are related to a recent event [203].

Following the big picture just presented, we identify two directions from which to access information in social media: (i) the people active within social media and (ii) their individual utterances. We refer to these “access directions” as entry points, since they act as a doorway to the information in social media. We now zoom in on these entry points and how they shape the research in this thesis. The main aim that we want to address is to improve searching for people and their utterances in social media to offer intelligent access to information in those media. We address this overarching goal by tackling a series
1. Introduction

of smaller research questions and aggregate the results in Chapter 9, the conclusions.

We start by exploring how people search for people. It is estimated that 11–17% of web queries contain a person name, and, more so, 4% of web queries are person name queries only [7]. No fewer than 57% of adult Internet users use a search engine to search for their own name [122]. The goal of the searcher is to get more information about the person for whom she is looking, for example in the form of online profiles, pictures, or news articles. In this part of the research, we do not look at the utterances of people we are looking for, but we analyze the query logs of a people search engine to gain insight in search behavior, much like previous work in web search [26], blog search [138], and scientific literature search [90]. We also explore the relation between social media and search behavior and wonder, for example, if social media influence which persons users are looking for? And on the result side, are users mostly interested in results from social media platforms or is it other information they would like to see? We ask:

RQ 1 How do users go about searching for people, when offered a specialized people search engine to access these people’s profiles?

1. What are the general usage statistics of a people search engine and how do these compare to general web search engines?
2. Can we identify different types of person name queries that users issue to the search engine?
3. Is automatic classification of queries into the different types feasible? What kind of features are most useful for this task?
4. Can we indicate where the interest in certain queries (e.g., popular names) comes from? And what do users want to see as results?
5. On a higher level of aggregation, can we identify different types of session (i.e., a set of queries from one user) and returning users?
6. Can we identify future research directions based on (unexpected) findings in the query logs?

So far, we have ignored people’s utterances. In our next research question we bring these in. We can represent a person by her utterances, and use these utterances to get an idea of what this person’s main interests are. Using this information, we can, for a given topic, suggest people who are interested in, or knowledgeable about that topic. For this set of questions, we focus on blogs as our social media platform and use bloggers’ posts to find bloggers we are interested in. Of interest here is the way we represent the blogger and how we aggregate information from an individual utterance level to a person level. For the latter questions we identify three types of model: (i) post-based models construct a post ranking and aggregate scores of individual posts to a blogger score; (ii) blog-based models create a representation of the entire blog and use this for ranking bloggers; (iii) a two-stage model that exploits the following observation about human strategies for identifying complex information objects (e.g., blogs, people, . . .). Prior to in-depth examination of complex information objects, humans display exploratory search behavior triggered by salient features of such objects [98]. We translate this strategy to a blogger finding model and we ask:
RQ 2 Can we effectively and efficiently search for people who show a recurring interest in a topic using an index of utterances?

1. Can we model the task of blogger finding as an association finding task?
2. How do our implementations of the post-based (Posting) and blog-based (Blogger) models compare to each other in terms of retrieval effectiveness and efficiency?
3. Can we introduce different association strength indicators between posts and blogger and how do they influence performance?
4. Can we combine the strengths of the two models and how does this new, two-stage model perform compared to our baselines?
5. Can we improve efficiency by limiting the number of posts we look at or by reducing the document representations (e.g., title-only)?

We move away from people as the unit of retrieval and dive into the area of finding relevant utterances. Here, we start by looking at characteristics of the utterances themselves and touch on the people who produced them. Without rules and editors in social media platforms people can write whatever they want, in whatever form they feel like. However, when looking for relevant information on a topic, we expect people to prefer utterances that have a certain level of quality, and that they “believe” more than other utterances. We refer to these aspects of information as “credibility.” The notion of credibility has been substantiated for the blogosphere by Rubin and Liddy [160], who proposed a credibility framework for blogs. Credibility is a concept that can apply at the level of users and at the level of their individual utterances. We ask:

RQ 3 Can we use the notion of credibility of utterances and people to improve on the task of retrieving relevant blog posts?

1. Given the credibility framework developed in [160], which indicators can we measure from the text of blog posts?
2. Can we incorporate credibility-inspired indicators in the retrieval process, keeping in mind the precision-oriented nature of the task? We try two methods: (i) “Credibility-inspired reranking” based on credibility-inspired scores and (ii) “Combined reranking” based on credibility-inspired scores and retrieval scores.
3. Can individual credibility-inspired indicators improve precision over a strong baseline?
4. Can we improve performance (further) by combining indicators in blog and post-level groups? And by combining them all?

One of the grand challenges in most retrieval tasks is to bridge the vocabulary gap between a user and her information need on the one hand and the relevant documents on the other [11]. An often-used technique to overcome this challenge is pseudo-relevance
feedback, where the original query is expanded using terms from the top ranked documents [126]. Given the noisy character of social media utterances, it is difficult to improve effectiveness using pseudo-relevance feedback [6, 82]. To counter the noisiness of the data in social media, we use the fact that people are part of a real-world environment and that this environment influences their utterances. We incorporate information from the environment in query expansion, resulting in external query expansion (i.e., query expansion using external sources) [44]. We aim at overcoming the problems that result from very noisy data. We ask:

**RQ 4** Can we incorporate information from the environment, like news or general knowledge, in finding blog posts using external expansion?

1. Can we effectively apply external expansion in the retrieval of blog posts?
2. Does conditioning the external collection on the query help improve retrieval performance?
3. Which of the external collections is most beneficial for query expansion in blog post retrieval?
4. Does our model show similar behavior across topics or do we observe strong per-topic differences?

Finally, we observe that utterances are not isolated. Unlike the preceding research question, in which we explore the environment that influences what a person writes about, here, we focus on the immediate environment in which utterances are produced. In many social media platforms this immediate environment is very structured, such as “blog–blog post–comments” and “forum–thread–post–quote,” creating various levels of context. We believe the information contained in (nearby) context levels within the same platform can be used to find relevant utterances, as the context provides additional evidence of relevance for these utterances. The work on these context levels is related to the incorporation of the environment done in the previous research questions. Besides the context levels, we also take the notion of credibility from blogs and translate it to another social media platform. Here, we focus on mailing lists, which record the conversations of a virtual community drawn together by a shared task or by a common interest [142]. In the end, we ask:

**RQ 5** Can we incorporate information from the utterances’ contexts in the task of finding emails?

1. Can we use the various context levels of an email archive levels to improve performance on finding relevant emails?
2. Which of these context levels is most beneficial for retrieval performance?
3. Can we further improve email search using credibility-inspired indicators as introduced in Chapter 6?

In each of the research chapters (Chapters 4–8) we seek answers to the research questions stated above. The answers are given in the conclusions of each chapter and are summarized in Chapter 9 of this thesis. In the next sections we list the contributions that this thesis makes to the field and we give an overview of the thesis and its origins.
1.3 Main Contributions

The main contributions of this thesis are listed below.

- **Insight in search behavior for a people search engine** – We analyze search behavior of users of a people search engine and offer insights in general usage statistics and the result types they most often click on. We give recommendations for people search based on observations from the query logs.

- **Classification scheme for people search** – We propose a classification scheme for people queries and evaluate automatic classification of queries into these classes. We also propose classification schemes for sessions of people queries and users of a people search engine.

- **The relation between people search and social media** – We present a case study that indicates how social media, traditional media, and people search activity are related.

- **Efficient and effective models for blogger finding** – We present three blogger finding models, each with their own pros and cons. We show how the models perform, both from an effectiveness and an efficiency perspective.

- **Measurable credibility-inspired indicators for social media utterances** – Based on a previously defined credibility framework we offer translations of items in this framework to measurable credibility-inspired indicators for blogs. We propose two ways of using the credibility-inspired indicators in a retrieval task.

- **A general model for external query expansion** – We propose a new general external query expansion model, that uses evidence from external collections to arrive at a better query representation. The main feature of the model, taking into account the query-dependent collection importance, is thoroughly analyzed and compared to previous approaches. We also analyze the performance of various external collections as sources for query expansion.

- **Methods to incorporate the structured environments in email search** – We propose a way of using the immediate context levels in email finding, much like the external collections. We also translate the blog credibility indicators to the domain of email search and analyze their performance.

1.4 Thesis Overview

Besides the current chapter, the thesis consists of two chapters covering the prerequisites and methodology, five research chapters containing our core contributions plus a concluding chapter:

**Chapter 2 - Background:** Here, we present a general introduction to information retrieval and various retrieval models. Each of the research chapters has its own related work section, in which we focus on query log analyses and retrieval in social media.
1. Introduction

Chapter 3 - Experimental Methodology: We provide details on experimental settings that recur in various chapters of this thesis. Amongst others, we discuss document collections, topic sets, and evaluation metrics. We provide details on our baseline retrieval model (language modeling for IR), which recurs in Chapters 5–8.

Chapter 4 - Searching for People: The first of five research chapters introduces the task of people search. Given a person name query, return information about this person (e.g., social media profiles, news articles, . . .). We analyze query logs of a people search engine and provide insights in the general search behavior for this search engine. On top of that, we introduce three person query types and explore sessions and users of this type of search engine. Observations made in this chapter serve as input to Chapter 7 and lead to a set of recommendations for people search.

Chapter 5 - Finding Bloggers: In this chapter we propose three models for finding bloggers that show a recurring interest in a given topic. Unlike Chapter 4 we use a blogger’s utterances for this task and explore how we can use information about individual blog posts in the task of blogger finding. We explore both effectiveness and efficiency of the proposed models, and analyze the results on a per topic basis.

Chapter 6 - Credibility-Inspired Ranking for Blog Post Retrieval: Based on a previously introduced credibility framework for blogs, we introduce credibility-inspired indicators on the user and utterance level that we can estimate from textual information. We incorporate these indicators in the task of blog post retrieval in two ways and analyze the impact of the indicators on the performance on this task.

Chapter 7 - Exploiting the Environment in Blog Post Retrieval: Exploiting the environment for blog post retrieval can be done through query expansion on external document collections. We propose a generative blog post retrieval model that uses information from external sources and we show how making the choice of external collection dependent on the query is beneficial. We compare results to a previously proposed mixture of external collections that ignores query-dependent collection importance.

Chapter 8 - Using Contextual Information for Email Finding: Here, we take ideas from Chapters 6 and 7 and translate them to the setting of email finding. First, we explore how an email’s direct context can be used to improve its retrievability. We show how using the various context levels in a mailing list (e.g., threads, community, . . .) can improve on email finding and analyze the portability of credibility-inspired indicators to a different social media platform.

Chapter 9 - Conclusions: We go back to the research questions introduced in this chapter and provide their answers. Finally, we discuss future directions of research.

Chapter 2 serves as background to the research in the technical chapters and can be read if additional insight in the field is required. Chapter 3 provides necessary information on the test collections and evaluation metrics that are used in the technical chapters and gives additional details on the baseline retrieval model. Each of the research Chapters 4 to 8 can be read individually, as the contents of these chapters is not dependent on other
1.5 Origins

The work presented in this thesis is based on a number of papers, of which details can be found in the bibliography. The analysis presented in Chapter 4 was first presented in [203] and additional analysis and experiments were published in [22]. The blog feed search models in Chapter 5 were introduced in [16, 201] and further built upon in [197, 202]. The work on credibility-inspired ranking in Chapter 6 was first published in [194] and expanded in [196]. The work in Chapter 7 is based on material published in [198], with additional insights published in [204]. Finally, the models for email search in Chapter 8 were presented in [199]. Other publication sources for this thesis include [17, 66, 82, 83, 128–130, 185, 189, 195, 200].