

Chapter 10

Conclusions and Future Directions



Today, the importance of entities has been broadly recognized and entities have become first-class citizens in many information access systems, including web, mobile, and enterprise search; question answering; and personal digital assistants. Entities have also become a meeting point for several research communities, including that of information retrieval, natural language processing, databases, and the Semantic Web. Many of the methods and tools we have described in this book, such as ranking entities, recognizing and linking entity mentions in documents and queries, or displaying entity cards, are now integral components of modern search systems.

Is this the end of the road? Certainly not. It would be going too far to label those core tasks, like entity ranking and entity linking, as “solved.” Obviously, there is still (plenty of) room for improvement. Also, it is not yet clear which techniques will be the “BM25’s” of the entity world, as stable and reliable solutions. Only time will tell. Nevertheless, we have reached a point where these methods are “good enough” to be used as basic building blocks in more complex systems. Perhaps it is time to look beyond these core tasks. As we are approaching the end of this book, we shall attempt to look into the future and gauge what lies ahead. Many of the things we will discuss here have already begun to happen, while some other elements, or their exact form, are more of a speculation.

In Sect. 10.1, we shall summarize our progress so far. Where are we now and how did we get here? Then, in Sects. 10.2 and 10.3 we will attempt to look ahead and discuss some anticipated future developments. We will conclude with some final remarks in Sect. 10.4.

10.1 Summary of Progress

Let us take a step back and distill the progress achieved over the past years, organized around three main thematic areas. We shall also briefly mention open issues; we will elaborate on some of these in more detail in Sect. 10.2.

10.1.1 *Data*

We start by discussing data, as developments in the data landscape have been instrumental to the progress made thus far. Specifically, the availability of large-scale knowledge bases has played a key role in transforming the search experience. Many information access applications utilize knowledge bases as a rich, structured repository of entities, and, to a lesser extent, for ontological background knowledge. Knowledge about particular entities may be used to complement the traditional (document-oriented) search results, allow for direct answers and various knowledge panels, and facilitate content exploration and discovery. Knowledge bases also enable machine understanding of natural language text, by using entities as a pivot to connect unstructured and structured data sources (Chap. 5). In turn, massive volumes of unstructured documents may be utilized to populate KBs with additional entities and their properties (Chap. 6).

Open Issues Knowledge bases are inherently incomplete and keeping them up-to-date requires a continuous effort. Automatic knowledge acquisition is an active area of research. Open challenges include the discovery of long tail and emerging entities, and the quality of data (correctness and trustworthiness of facts); see Sect. 10.3.3 for further data-related issues.

10.1.2 *Retrieval Methods*

A significant portion of the book has been devoted to entity retrieval methods. Early approaches build on document retrieval techniques and focus on how to adopt those for various types of data, from unstructured to structured (Chap. 3). More recent approaches utilize the rich structure associated with entities in knowledge bases (Chap. 4). Many—in fact, most—of the other tasks we have addressed in this book were also cast as ranking problems, for instance, disambiguating entities that may refer to a particular mention in text (Sect. 5.6), filtering documents that contain vital information about an entity (Sect. 6.2), identifying target types of a query (Sect. 7.2), finding interpretations of a query (Sect. 7.3.4), or determining which facts to display on an entity card (Sect. 9.2.2). For all these tasks, the current state of the art involves a discriminative learning approach, i.e., learning-to-rank, employing a rich set of carefully designed features.

Open Issues It appears to be a “safe” recipe to tackle any ranking problem by hand-crafting a large set of features, then throwing machine learning at it. Indeed, the importance of feature engineering is not to be underestimated. Nevertheless, one might argue that this general approach can even get rather mechanical, and scientifically less interesting, after a while. Neural methods, especially deep learning, hold the promise of learning directly from raw data, without such labor-intensive feature engineering. Extending traditional IR models to incorporate word embeddings has already proven effective for various entity-related search tasks, see, e.g., [5, 7, 13, 16, 17, 19]. Developing end-to-end architectures, which more fully embrace neural modeling, is an exciting and active research direction [12, 21]. Yet, it remains to be seen if deep learning can categorically outclass other approaches, and whether it will surpass all other forms of machine learning and take over the entire field of IR (as it did with computer vision, speech recognition, and machine translation). Even if it does, one might say that all this means is that feature engineering will get replaced by network engineering. Another issue here will surely be the availability of training data. In that regard, industry has a distinct advantage over academia, as target relevance labels may be derived on a much larger scale from usage data.

While the core entity-oriented retrieval tasks described above both merit and have the potential for further improvement, another open issue is how to combine these into more complex useful applications. After all, our eventual goal should be aiding users in achieving their goals, i.e., completing their tasks, which goes far beyond the ranking of items; see Sect. 10.3.2.

10.1.3 *Understanding and Interacting with Users*

Users increasingly expect search engines to understand them and respond to their information needs more directly than just serving documents matching the query terms. Today, the search box functions more like a “request box,” and queries are answered by rich search result pages, including direct answers and interactive widgets (maps, currency conversion, etc.). We have looked at how to utilize entities and types to understand information needs (Chap. 7) and to provide an enhanced search experience (Chap. 9).

Open Issues Search has become a consumer experience. Major search engines are continuously introducing new types of “functional” results (interactive widgets), enabling users to do more and more, without leaving the SERP. Result presentation and interacting with entities still offer plenty of opportunities for research and innovation. One recent line of work focuses on actionable knowledge bases, i.e., identifying potential actions that can be performed on a given entity [4].

Another open issue in this area is that search (or, more broadly, information access) is moving from desktop to mobile and from text to voice. Personal digital assistants are increasingly being used to respond to natural language questions. We can say that search is becoming a conversation between humans and machines; see Sect. 10.3.1.

10.2 A Peek into the Future

In this section, we present a fictional conversation that takes place sometime in the not-too-distant future, between a user, say a male university professor, and an intelligent personal assistant, simply referred to as “AI.” This conversation could in fact happen on any device, but for the sake of illustrating certain points, we shall assume that the device is a mobile phone and that the user interacts with the device via spoken natural language. The conversation, which will later be referred to as *scenario*, is accompanied by some narrative.

AI I see you're wasting time away on Facebook. Do you have time now to talk about your holiday plans?

The first thing to notice is that it is the AI that initiates the conversation. Based on the user's current activity and past behavioral patterns, it decides that now would be a good time to address a future information need.

Sure. I want an active holiday with the family in beautiful nature.

AI It sounds like you would definitely love Norway. A cabin in the mountains maybe?


Could be. But I want to go kayaking and also catch some fish. And not too much rain, please.

AI And something fun for the kids to do nearby, I suppose?

Of course.

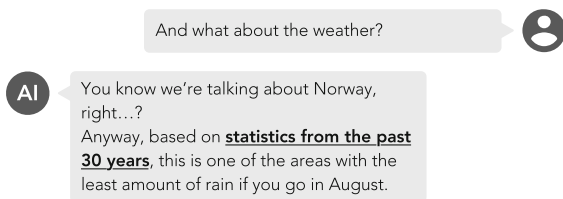
The AI refines the requirements iteratively by asking a series of questions. Observe that it has knowledge of the user's background (family situation). After having the initial requirements clarified, it comes up with a specific suggestion:

AI How does **Oltedal** sound? People visiting there have been quite successful with catching lake trout, based on what I found on Instagram.

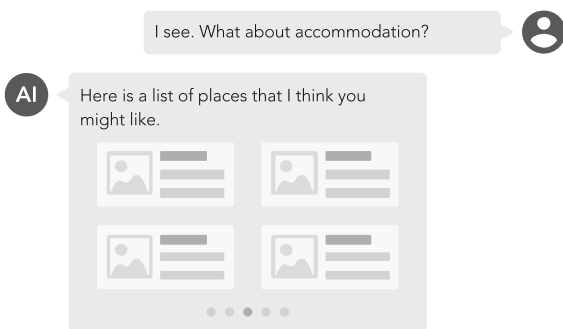


There is also a **theme park** and **horseback riding**, both within 50 kms.

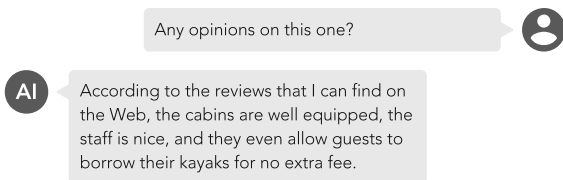
Links are boldfaced and underlined in the response text; the user could follow these for more information. The photos are also clickable.



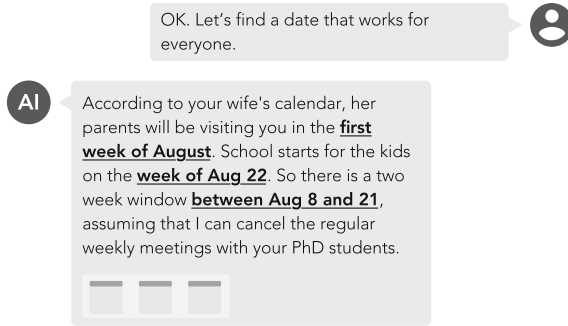
Two things are worth pointing out in the AI's response. One is that it has a (certain) sense of humor. Humor is an essential human communication behavior. The other is that it is able to give a compact answer to the question, and backs it up with a link to the source (evidence) that the answer was based on.



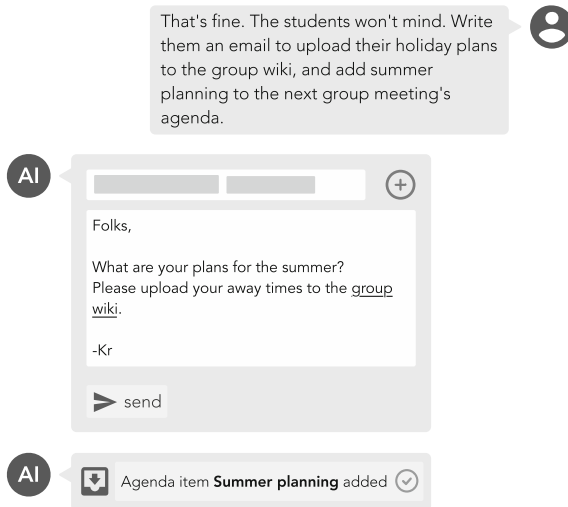
The AI presents a list of (personalized) accommodation options. After examining the results, the user selects one of the items and asks for further information.



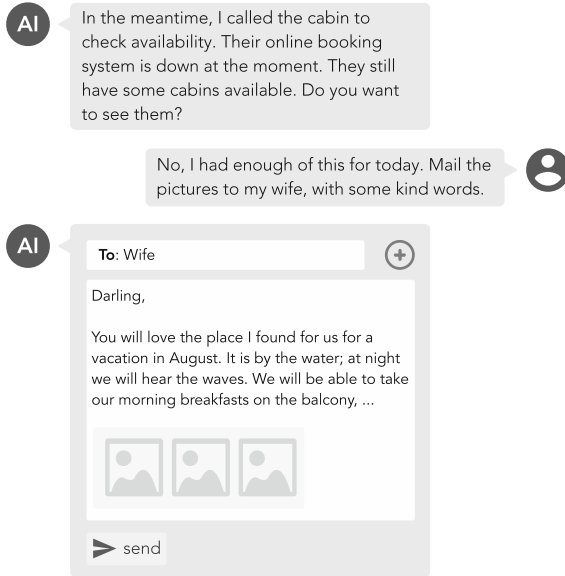
With this reply, the AI demonstrates some impressive summarization skills. It focuses on aspects that are likely of interest to the user.



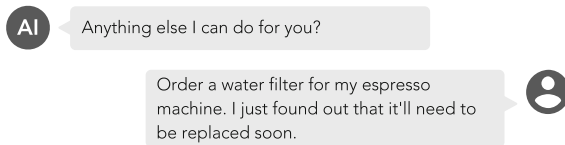
At this point, we are moving away from what was a kind of exploratory search scenario to a different type of information access problem, where the AI helps to automatically manage scheduling.



First, the AI prepares a draft of the requested email message in the user's style. The user can refine the text before sending. Then, the AI adds the agenda item to the team wiki and displays a success notification received from the wiki software.



The AI attempts to check availability via an online booking system. Realizing that it is currently not accessible, it decides to resort to more conventional means of communication and calls the place over the phone. Next, it composes an email message, combining the user’s language model with a flavor of “kind.” After reviewing—and perhaps correcting the AI’s over-the-top romantic vibe here-and-there—the user decides to send off the message.



With this task completed, the AI asks if it could be of any more assistance. Then, it receives a request of a different nature.

10.3 Future Research Directions

Below, we discuss a number of directions and areas for future research. Along the way, we will occasionally make references to certain elements of the scenario presented in the previous section.

10.3.1 *Understanding and Interacting with Users*

Search Is a Conversation For many years, keyword queries have been the *lingua franca* of information access. This, however, is changing. With the emergence of the mobile-over-desktop culture of information consumption, and the advancement in voice recognition technologies, voice search is gaining ground. In 2015, Google reported that the volume of mobile search has surpassed that of desktop search in several countries.¹ As of 2016, around 20% of queries on Google mobile devices are voice input in the USA [14]. Voice queries are not only longer on average than text queries but also use richer language [9]. But there is more. Voice search facilitates the possibility of a speech dialogue with the user. Such natural language interfaces are already a reality, as manifested in personal digital assistants, such as the Google Assistant, Apple’s Siri, Microsoft’s Cortana, or Amazon’s Alexa.

Conversational search offers many possibilities, such as the ability to ask the user for clarification, if needed. It also presents many challenges, as the system no longer returns massive search engine result pages, somewhere on which the user hopefully finds what she was looking for. The response needs to be more “intelligent,” i.e., comprehensive and spot-on. In this regard, voice-based result presentation that enables a completely hand-free interaction with the user still has a long way to go [9]. There is also a need for novel evaluation measures that can capture user satisfaction in a conversational setting. A good conversation entails more than just the fulfillment of an information need; among others, it should flow and be engaging, be just about the right length, and, occasionally, even humorous.

Anticipating Information Needs The traditional way of information access is *reactive*: The system responds to a user-issued request. A *proactive* system, on the other hand, “would anticipate and address the user’s information need, without requiring the user to issue (type or speak) a query” [3]. Hence, this paradigm is also known as *zero-query search* [1]. Our scenario started out with the AI proactively bringing up a future information need. We observed proactive recommendations in later parts of the conversation too, when considering additional criteria in exploratory search (activities for kids) and when figuring out how to schedule the vacation (cancelling meetings). Some of today’s personal digital assistants are already capable of pre-fetching information cards based on users’ behavioral patterns or upcoming events (e.g., Google Now and Microsoft Cortana). Recent research has focused on a number of specific problems in this space, including modeling user interests [20], predicting when users will perform a repetitive task again in the future [15], identifying what information needs people have in a given context [3], and determining the right context for pushing proactive recommendations [6]. With intelligent devices capable of sensing the user’s environment (location, and even pulse rate or blood pressure using wearable devices), there are increasingly more contextual signals that may be utilized. Notably, current work

¹<https://adwords.googleblog.com/2015/05/building-for-next-moment.html>.

is limited to near-term information needs. The area of anticipating more long-term information needs (such as reminding a user months in advance about planning a vacation or finding a school for a child that is going to go to school next year) has not been explored yet.

Verification and Explainability As we move away from ranked lists of items to direct answers and summaries, it becomes crucial to allow for the verification of the system's responses. What is the right form of explanation? In many cases, providing access to the raw data is sufficient. We have seen several examples of this in our scenario, when the AI provided links to pages about weather statistics, reviews, and calendars. In other cases, it may not be possible to refer to a single source; then, the user should be granted access to some intermediate data representation. It is an open issue how to make those intermediate representations suitable for human consumption. These questions also relate to the broader problem area of providing explanations of algorithmic decisions that significantly affect an individual (particularly legally or financially), which is to be a human right according to the European Union General Data Protection Regulation ("right to explanation"), to take effect in 2018 [8].

Personalization In our scenario, we could observe a high degree of personalization, including the interaction with the user, the generation of responses, and the language usage when executing tasks on the user's behalf. Personal digital assistants are expected to deliver such a personalized user experience. To be able to do that, they will need to get to "know" the user, her habits, preferences, and the things she cares about. With human assistants, there is often a more-or-less clear separation between work and private matters. This is not the case with digital assistants; most users would likely use the same personal AI for any and all kinds of business they encounter. This brings up many issues around trust, privacy, and data protection. Digital assistants must be aware of the user's momentarily situation and context too.

10.3.2 *Complex Information Needs and Task Completion*

Major web search engines have made a great progress with answering one-shot queries with rich search result pages, thereby putting the bar rather high regarding the search experience. Users now expect intelligent personal assistants to respond with direct answers as opposed to a ranked list of results. Thus, it may be fitting to refer to these systems no longer as search engines but as *answering engines*. Intelligent agents are further capable of assisting users in "getting things done," such as making calendar appointments or setting reminders. However, neither web search engines nor digital assistants have the capability yet to handle truly complex tasks, such as the holiday planning in our scenario. These complex information needs require a better understanding and modeling of the user's high-level goals. It requires no less than a paradigm shift, from answering engines to *task-completion engines* [2]. Entities will continue to play a key role here, for modeling users, tasks, and context.

10.3.3 *Data and Knowledge*

On-the-Fly Information Extraction Despite all automatic knowledge acquisition efforts, there will always be long-tail entities that are not contained in any knowledge base. Moreover, even if the entities in question are present in a knowledge base, it is not possible to capture all information associated with them, due to the finite vocabulary of knowledge base predicates. Consequently, we will continue to come across information needs to which the answer is “out there” in some digital form, but not yet contained in a knowledge base. For example, in our scenario, this could be the case with some accommodations at obscure locations. These situations may be handled by on-the-fly information extraction techniques.

Personal Knowledge Base In our scenario, the user has made numerous references to entities he was in some way related to: “my kids,” “my wife,” “my group,” “my espresso machine,” etc. These entities constitute the users’ *personal knowledge base*, i.e., the universe of things he cares about. Throughout interactions with the user, the entities of this universe may be mapped onto the same data representation model that knowledge bases use. It is also possible to make “same-as” links to other knowledge repositories that contain the same entity (e.g., the espresso machine). Some entities, however, will reside only in the user’s personal KB. What is powerful about this idea is that the same methods and techniques we have discussed for general-purpose KBs are readily applicable to a personal KB. The problem thus boils down to the automatic population and maintenance of the personal KB.

Commonsense Knowledge Knowledge bases have largely focused on accumulating factual knowledge about specific entities. An intelligent system, such as a personal digital assistant, however, needs a much broader understanding of the world. Simple statements like “things fall down, not up” and “open the door before entering” are obvious to humans but not to machines. To endow computers with common sense is one of the long-standing goals of AI research. Some projects, such as Cyc [10] or ConceptNet [11], have begun to amass large collections of such commonsense knowledge. However, “there is still a long way to go for computers to learn what every child knows” [18].

10.4 Concluding Remarks

Reaching the end of this book, it may be appropriate to have a moment of reflection. Information technology has changed and will continue to change our lives. We are increasingly more surrounded by intelligent autonomous systems (which we like to call AI): personal assistants, self-driving cars, smart homes, etc. There are some thought-provoking open questions here related to responsibility: If a fatal accident happens involving an autonomous vehicle or a disastrous decision is made based on false information served by a search engine (which perhaps retrieved

it from some underlying knowledge base), who is responsible for that? Surely, the company behind the given product should take some responsibility. But then, would it ultimately come down to the individual software engineer who wrote the corresponding piece of code (or to the knowledge engineer who was responsible for that entry ending up in the KB)? Or would the blame be put on the end user, who did not study or consider carefully enough the terms of usage? These are important and challenging regulatory issues on which conversations have already started.

We are now in the third AI spring, which draws mixed reactions from people: great excitement, overblown expectations because of the hype, and fear. Technological singularity, i.e., the emergence of an (evil) artificial superintelligence that would cause the human race to go extinct—in the author’s opinion—is merely a dystopia that Hollywood loves to portray in speculative fiction. Technology itself is not good or evil—it depends on how we use it. It appears though that as time goes on, increasingly more technology will be “forced” on us. Yet, we have the free will and responsibility to decide what technology we want to use or adopt. Importantly, technology should enable and not distract us on that awesome journey, with its ups and downs, that is called Life. Along the way, we should take the time to contemplate on the deeper questions of existence, being, and *identity*—searching for the answers to those questions is what it means to be a human. No computer system, however intelligent, will ever be able to do that for us.

References

1. Allan, J., Croft, B., Moffat, A., Sanderson, M.: Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the Second Strategic Workshop on information retrieval in IJIR. *SIGIR Forum* **46**(1), 2–32 (2012)
2. Balog, K.: Task-completion engines: A vision with a plan. In: Proceedings of the First International Workshop on Supporting Complex Search Tasks, SCST ’15 (2015)
3. Benetka, J.R., Balog, K., Nørsvåg, K.: Anticipating information needs based on check-in activity. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining, WSDM ’17, pp. 41–50. ACM (2017). doi: [10.1145/3018661.3018679](https://doi.org/10.1145/3018661.3018679)
4. Blanco, R., Joho, H., Jatowt, A., Yu, H., Yamamoto, S.: NTCIR Actionable Knowledge Graph task (2017)
5. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM ’15, pp. 179–188. ACM (2015). doi: [10.1145/2684822.2685317](https://doi.org/10.1145/2684822.2685317)
6. Braunhofer, M., Ricci, F., Lamche, B., Wörndl, W.: A context-aware model for proactive recommender systems in the tourism domain. In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, MobileHCI ’15, pp. 1070–1075. ACM (2015). doi: [10.1145/2786567.2794332](https://doi.org/10.1145/2786567.2794332)
7. Garigliotti, D., Hasibi, F., Balog, K.: Target type identification for entity-bearing queries. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’17. ACM (2017). doi: [10.1145/3077136.3080659](https://doi.org/10.1145/3077136.3080659)
8. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a “right to explanation”. ArXiv e-prints (2016)
9. Guy, I.: Searching by talking: Analysis of voice queries on mobile web search. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’16, pp. 35–44. ACM (2016). doi: [10.1145/2911451.2911525](https://doi.org/10.1145/2911451.2911525)

10. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM* **38**(11), 33–38 (1995). doi: [10.1145/219717.219745](https://doi.org/10.1145/219717.219745)
11. Liu, H., Singh, P.: ConceptNet - A practical commonsense reasoning tool-kit. *BT Technology Journal* **22**(4), 211–226 (2004). doi: [10.1023/B:BTTJ.0000047600.45421.6d](https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d)
12. Mitra, B., Craswell, N.: Neural models for information retrieval. ArXiv e-prints (2017)
13. Pappu, A., Blanco, R., Mehdad, Y., Stent, A., Thadani, K.: Lightweight multilingual entity extraction and linking. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17, pp. 365–374. ACM (2017). doi: [10.1145/3018661.3018724](https://doi.org/10.1145/3018661.3018724)
14. Pichai, S.: Google I/O 2016 keynote (2016)
15. Song, Y., Guo, Q.: Query-less: Predicting task repetition for nextgen proactive search and recommendation engines. In: Proceedings of the 25th International Conference on World Wide Web, WWW '16, pp. 543–553 (2016). doi: [10.1145/2872427.2883020](https://doi.org/10.1145/2872427.2883020)
16. Van Gysel, C., de Rijke, M., Kanoulas, E.: Learning latent vector spaces for product search. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM '16, pp. 165–174. ACM (2016a). doi: [10.1145/2983323.2983702](https://doi.org/10.1145/2983323.2983702)
17. Van Gysel, C., de Rijke, M., Worring, M.: Unsupervised, efficient and semantic expertise retrieval. In: Proceedings of the 25th International Conference on World Wide Web, WWW '16, pp. 1069–1079 (2016b). doi: [10.1145/2872427.2882974](https://doi.org/10.1145/2872427.2882974)
18. Weikum, G., Hoffart, J., Suchanek, F.: Ten years of knowledge harvesting: Lessons and challenges. *IEEE Data Eng. Bull.* **39**(3), 41–50 (2016)
19. Xiong, C., Power, R., Callan, J.: Explicit semantic ranking for academic search via knowledge graph embedding. In: Proceedings of the 26th International Conference on World Wide Web, WWW '17, pp. 1271–1279. International World Wide Web Conferences Steering Committee (2017). doi: [10.1145/3038912.3052558](https://doi.org/10.1145/3038912.3052558)
20. Yang, L., Guo, Q., Song, Y., Meng, S., Shokouhi, M., McDonald, K., Croft, W.B.: Modeling user interests for zero-query ranking. In: Proceedings of the 38th European Conference on IR Research, ECIR '16, pp. 171–184 (2016). doi: [10.1007/978-3-319-30671-1_13](https://doi.org/10.1007/978-3-319-30671-1_13)
21. Zhang, Y., Mustafizur Rahman, M., Braylan, A., Dang, B., Chang, H.L., Kim, H., McNamara, Q., Angert, A., Banner, E., Khetan, V., McDonnell, T., Thanh Nguyen, A., Xu, D., Wallace, B., Lease, M.: Neural information retrieval: A literature review. ArXiv e-prints (2016)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

