An Introduction to Neural Information Retrieval
Other titles in Foundations and Trends® in Information Retrieval

*Efficient Query Processing for Scalable Web Search*
Nicola Tonellotto, Craig Macdonald and Iadh Ounis

*Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text*
Ross S. Purves, Paul Clough, Christopher B. Jones, Mark H. Hall and Vanessa Murdock

*Web Forum Retrieval and Text Analytics: A Survey*
Doris Hoogeveen, Li Wang, Timothy Baldwin and Karin M. Verspoor
An Introduction to Neural Information Retrieval

Bhaskar Mitra
Microsoft, University College London
Montreal, Canada
bmitra@microsoft.com

Nick Craswell
Microsoft
Bellevue, USA
nickcr@microsoft.com

Full text available at: http://dx.doi.org/10.1561/1500000061
Editorial Scope

Topics

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2019, Volume 13, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.
Contents

1 Introduction 2

2 Fundamentals of text retrieval 8
   2.1 IR tasks .................................................. 8
   2.2 Desiderata of IR models ................................. 11
   2.3 Notation .................................................. 17
   2.4 Metrics ................................................... 17
   2.5 Traditional IR models .................................... 19
   2.6 Neural approaches to IR ................................ 23

3 Unsupervised learning of term representations 27
   3.1 A tale of two representations ............................ 27
   3.2 Notions of similarity ..................................... 31
   3.3 Observed feature spaces .................................. 33
   3.4 Latent feature spaces .................................... 36

4 Term embeddings for IR 44
   4.1 Query-document matching .................................. 46
   4.2 Query expansion ........................................... 52

5 Supervised learning to rank 55
   5.1 Input features ............................................... 56
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>Loss functions</td>
<td>57</td>
</tr>
<tr>
<td>6</td>
<td>Deep neural networks</td>
<td>68</td>
</tr>
<tr>
<td>6.1</td>
<td>Input text representations</td>
<td>70</td>
</tr>
<tr>
<td>6.2</td>
<td>Standard architectures</td>
<td>71</td>
</tr>
<tr>
<td>6.3</td>
<td>Neural toolkits</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>Deep neural networks for IR</td>
<td>81</td>
</tr>
<tr>
<td>7.1</td>
<td>Document autoencoders</td>
<td>83</td>
</tr>
<tr>
<td>7.2</td>
<td>Siamese networks</td>
<td>84</td>
</tr>
<tr>
<td>7.3</td>
<td>Interaction-based networks</td>
<td>86</td>
</tr>
<tr>
<td>7.4</td>
<td>Lexical and semantic matching</td>
<td>87</td>
</tr>
<tr>
<td>7.5</td>
<td>Matching with multiple document fields</td>
<td>91</td>
</tr>
<tr>
<td>8</td>
<td>Conclusion</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Acknowledgements</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>99</td>
</tr>
</tbody>
</table>
ABSTRACT

Neural ranking models for information retrieval (IR) use shallow or deep neural networks to rank search results in response to a query. Traditional learning to rank models employ supervised machine learning (ML) techniques—including neural networks—over hand-crafted IR features. By contrast, more recently proposed neural models learn representations of language from raw text that can bridge the gap between query and document vocabulary. Unlike classical learning to rank models and non-neural approaches to IR, these new ML techniques are data-hungry, requiring large scale training data before they can be deployed. This tutorial introduces basic concepts and intuitions behind neural IR models, and places them in the context of classical non-neural approaches to IR. We begin by introducing fundamental concepts of retrieval and different neural and non-neural approaches to unsupervised learning of vector representations of text. We then review IR methods that employ these pre-trained neural vector representations without learning the IR task end-to-end. We introduce the Learning to Rank (LTR) framework next, discussing standard loss functions for ranking. We follow that with an overview of deep neural networks (DNNs), including standard architectures and implementations. Finally, we review supervised neural learning to rank models, including recent DNN architectures trained end-to-end for ranking tasks. We conclude with a discussion on potential future directions for neural IR.
Since the turn of the decade, there have been dramatic improvements in performance in computer vision, speech recognition, and machine translation tasks, witnessed in research and in real-world applications (LeCun et al., 2015). These breakthroughs were largely fuelled by recent advances in neural network models, usually with multiple hidden layers, known as deep architectures (Krizhevsky et al., 2012; LeCun et al., 2015; Hinton et al., 2012; Bahdanau et al., 2014; Deng, Yu, et al., 2014) combined with the availability of large datasets (Wissner-Gross, 2016) and cheap compute power for model training. Exciting novel applications, such as conversational agents (Vinyals and Le, 2015; Sordoni et al., 2015b), have also emerged, as well as game-playing agents with human-level performance (Silver et al., 2016; Mnih et al., 2015). Work has now begun in the information retrieval (IR) community to apply these neural methods, leading to the possibility of advancing the state of the art or even achieving breakthrough performance as in these other fields.

Retrieval of information can take many forms (White, 2016). Users can express their information need in the form of a text query—by typing on a keyboard, by selecting a query suggestion, or by voice recognition—or the query can be in the form of an image, or in some
cases the need can be implicit. Retrieval can involve ranking existing pieces of content, such as documents or short-text answers, or composing new responses incorporating retrieved information. Both the information need and the retrieved results may use the same modality (e.g., retrieving text documents in response to keyword queries), or be different (e.g., image search using text queries). If the query is ambiguous, retrieval system may consider user history, physical location, temporal changes in information, or other context when ranking results. IR systems may also help users formulate their intent (e.g., via query auto-completion or query suggestion) and can extract succinct summaries of results that take the user’s query into account.

We note that many natural language processing tasks exist that are not IR. Machine translation of text from one human language to another is not an IR task, because translating language and searching a corpus to satisfy a user’s information need are different. However, translation could be used in an IR system, to enable cross-language retrieval on a multilingual corpus (Oard and Diekema, 1998). Named entity linking, where text is disambiguated through linking to a knowledgebase, is not an IR task in itself. However, an IR system could use entity linking to enhance its performance on IR tasks. In general, many natural language processing tasks do not involve information access and retrieval, so are not IR tasks, but some can still be useful as part of a larger IR system.

Neural IR is the application of shallow or deep neural networks to IR tasks. Other natural language processing capabilities such as machine translation and named entity linking are not neural IR but could be used in an IR system.

Neural IR refers to the application of shallow or deep neural networks to retrieval tasks. Neural models have been employed in many
IR scenarios—including ad-hoc retrieval, recommender systems, multimedia search, and even conversational systems that generate answers in response to natural language questions. This tutorial serves as an introduction to neural methods for ranking documents in response to a query, an important IR task. We scope our discussions to a single task to allow for more thorough treatment of the fundamentals as opposed to providing a shallow survey of neural approaches to all IR tasks.

A search query may typically contain a few terms, while the document length, depending on the scenario, may range from a few terms to hundreds of sentences or more. Neural models for IR use vector representations of text, and usually contain a large number of parameters that need to be tuned. ML models with large set of parameters typically benefit from large quantity of training data (Brill, 2003; Taylor et al., 2006; Rajaraman, 2008; Halevy et al., 2009; Sun et al., 2017). Unlike traditional learning to rank (LTR) approaches (Liu, 2009) that train ML models over a set of hand-crafted features, recent neural models for IR typically accept the raw text of a query and document as input. Learning suitable representations of text also demands large-scale datasets for training (Mitra et al., 2017a). Therefore, unlike classical IR models, these neural approaches tend to be data hungry, with performance that improves with more training data.

Text representations can be learnt in an unsupervised or supervised fashion. The supervised approach uses IR data such as labelled query-document pairs, to learn a representation that is optimized end-to-end for the task at hand. If sufficient IR labels are not available, the unsupervised approach learns a representation using just the queries and/or documents. In the latter case, different unsupervised learning setups may lead to vector representations that capture different notions of text similarity. When applying such representations, the choice of unsupervised learning setup should be carefully considered, to yield a notion of text similarity that is suitable for the target task. Traditional IR models such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990) learn dense vector representations of terms and documents. Neural representation learning models share commonalities with these traditional approaches. Much of our understanding of these traditional
Figure 1.1: The percentage of neural IR papers at the ACM SIGIR conference—as determined by a manual inspection of the papers—shows a clear trend in the growing popularity of the field.

approaches from decades of research can be extended to these modern representation learning models.

In other fields, the design of neural network models has been informed by characteristics of the application and data. For example, the datasets and successful architectures are quite different in visual object recognition, speech recognition, and game playing agents. While IR shares some common attributes with the field of natural language processing, it also comes with its own set of unique challenges. IR systems must deal with short queries that may contain previously unseen vocabulary, to match against documents that vary in length, to find relevant documents that may also contain large sections of irrelevant text. IR systems should learn patterns in query and document text that indicate relevance, even if query and document use different vocabulary, and even if the patterns are task-specific or context-specific.

The goal of this tutorial is to introduce the fundamentals of neural IR, in context of traditional IR research, with visual examples to illustrate key concepts and a consistent mathematical notation for describing key models. Section 2 presents a survey of IR tasks, challenges, metrics
and non-neural models—as well as a brief overview of different neural approaches to IR. Section 3 introduces neural and non-neural methods for learning term embeddings, without the use of supervision from IR labels, and with a focus on the notions of similarity. Section 4 surveys some specific approaches for incorporating such unsupervised embeddings in IR. Section 5 introduces supervised learning to rank models. Section 6 introduces the fundamentals of deep models—including standard architectures and toolkits—before §7 surveys some specific approaches for incorporating deep neural networks (DNNs) in IR. Section 8 is our discussion, including future work, and conclusion.

Motivation for this tutorial  Neural IR is an emerging field. Research publication in the area has been increasing (Figure 1.1), along with relevant workshops (Craswell et al., 2016a; Craswell et al., 2016b; Craswell et al., 2017; Craswell et al., 2018), tutorials (Li and Lu, n.d.; Mitra and Craswell, 2017; Kenter et al., 2017; Kenter et al., 2018a; Kenter et al., 2018b), and plenary talks (Manning, 2016; Craswell, 2017). Because this growth in interest is fairly recent, some researchers with IR expertise may be unfamiliar with neural models, and other researchers who have already worked with neural models may be unfamiliar with IR. The purpose of this tutorial is to bridge the gap, by describing the relevant IR concepts and neural methods in the current literature.

A thorough review of the fundamentals of IR or neural networks, however, are beyond the scope of this tutorial. We refer interested readers to the following books for a general overview of traditional IR.

2. *Introduction to information retrieval*, by Manning et al. (2008).

Similarly, we recommend the following as companion reading materials for machine learning and neural network fundamentals.


References


Full text available at: http://dx.doi.org/10.1561/1500000061


De Saussure, F. 1916. “Cours de linguistique générale, publié par Ch”. De Saussure, F. 1916. “Cours de linguistique générale, publié par Ch”.


References


References


Full text available at: http://dx.doi.org/10.1561/1500000061


References


References


Full text available at: http://dx.doi.org/10.1561/1500000061


Mitra, B. and N. Craswell. 2015. “Query Auto-Completion for Rare Prefixes”. In: Proc. CIKM. ACM.


Rezende, D. J., S. Mohamed, and D. Wierstra. 2014. “Stochastic back-propagation and approximate inference in deep generative models”. In: Proc. ICML.
References


Full text available at: http://dx.doi.org/10.1561/1500000061


Full text available at: http://dx.doi.org/10.1561/1500000061
References


References


Full text available at: http://dx.doi.org/10.1561/1500000061


