Graphs and Information Retrieval

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen

op woensdag 22 maart 2023
om 12:00 uur precies

door

Chris Kamphuis
geboren op 22 maart 1993 te Oldenzaal, Nederland

Promotor:

prof. dr. ir. A.P. (Arjen) de Vries

Manuscriptcommissie:

Person A	(Affiliation)
Person B	(Affiliation)
Person C	(Affiliation)

This work is part of the research program Commit2Data with project number 628.011.001 (SQIREL-GRAPHS), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO).

Printed by Drukkerij

Typeset using LaTeX

ISBN: 111-11-11111-11-1

chriskamphuis.com

Contents

Chapter 1

Introduction

I also propose to consider the question, "Can machines think?" Instead of approaching this through a thought experiment like Turing did, nowadays one can approach this question by asking it to a search engine. When issuing this query to popular web search engines we get different results; the first result on Google is a passage generated from the article written by Turing, while the first result on Bing is a passage generated from a website that states machines can not think[1]. We use these systems that process queries every day in our lives to provide us information. Whereas Google and Bing are all purpose web engines that mainly focus on finding and retrieving web data, people also used specialized search systems in their day-to-day lives, examples are: Amazon / EBay for product search, NS for public transport in the Netherlands, Scholar / Zeta Alpha for scientific resources, Youtube / TikTok for Videos, or Facebook / LinkedIn for people. When searching for the query "Can machines think?", the approach of searching through text document only might be sufficient for the user. However in many cases when searching today, only considering text is not sufficient. When one wants to buy a product on Amazon, aspects other than text also need to be considered. Lets say for example you want to buy an IPhone; What is the price, which edition is the most recent, or which color does it have. When someone searches for people on LinkedIn, they are generally more interested in persons that have connections in common compared to complete strangers. If you are looking for

[1]However, if a machine can not think, can we trust the result presented by this algorithm?

someone to do a job, it is ideal that a shared connection can vouch for them.

1.1 Information Retrieval

Everything that is needed to process a query like, “Can machines think?”, is subject to research by the field of information retrieval.

1.1.1 Inverted Indexes

1.2 Relational Databases

Relational databases are usually used to store structure data.

1.3 Graphs

Instead of using columnar data, it might be more attractive to model your data using graphs.

Chapter 2

IR using Relational Databases

2.1 Introduction

Where commonly information retrieval researchers use inverted indexes as data structures, there is also a rich history of researchers using relational databases for representing the data in information retrieval systems.

In a more recent work [2] showed that the common used BM25 ranking function can also be easily expressed using relational tables. Their work specifically focused on the retrieval efficiency of several systems.

2.2 Reproducibility

We [1]

Chapter 3

From Tables to Graphs

Chapter 4

Data Modeling using Graphs

Chapter 5

Applications

Chapter 6

Conclusion

Bibliography

[1] Kamphuis, C., de Vries, A. P., Boytsov, L., and Lin, J. Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. In *Advances in Information Retrieval* (Cham, 2020), J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, and F. Martins, Eds., Springer International Publishing, pp. 28–34.

[2] Mühleisen, H., Samar, T., Lin, J., and de Vries, A. Old Dogs Are Great at New Tricks: Column Stores for IR Prototyping. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2014), SIGIR '14, Association for Computing Machinery, p. 863–866.

Summary

Samenvatting

Acknowlegements

Research Data Management

Curriculum Vitae