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 by Mühleisen et al. [4] 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.

The systems evaluated in this paper all purported to implement BM25, there was however a substantial difference between the effectiveness scores produced by these systems, as shown in table 2.1.

These results came out as quite a surprise as the authors took specific care to keep document pre-processing identical for all systems, but the observed difference in `MAP` of 3% absolute was the largest deviation in score reported.

2.2 Reproducibility

Not only did we observe the differences in effectiveness scores for BM25 in the paper by Mühleisen et al. [4]. In the SIGIR 2015 Work-

Table 2.1: Results presented by Mühleisen et al. [4]; MAP and P@5 on the ClueWeb12 collection are reported for five different systems that run BM25. As shown in the table, only the two database systems achieve the same effectiveness score. Both these systems were however developed by the same research group.

System	MAP	P@5
Indri	0.246	0.304
MonetDB & VectorWise	0.225	0.276
Lucene	0.216	0.265
Terrier	0.215	0.272

Table 2.2: Results from the RIGOR workshop[1], MAP@1000 on the .GOV2 collection is reported for four different systems that run BM25. As shown in the table, all four implementations report a different effectiveness score.

System	MAP@1000
ATIRE	0.2902
Lucene	0.3029
MG4J	0.2994
Terrier	0.2697

shop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR) [1] and the Open-Source IR Replicability Challenge (OSIRRC) workshop [2] similar results are observed. See tables 2.2 and 2.3 respectively.

It is not clear

2.3 Variants of BM25

Introduce what we did here

We examined several BM25 variants which will be introduced, how the variant varies from the original formulation as proposed by Robertson et al. is marked in red.

Table 2.3: Results from the OSIRRC workshop[2], AP, P@30, and NDCG@20 on the robust04 collection are reported for seven different systems that run BM25. As shown in the table, all implementations report (again) a different effectiveness score.

System	AP	P@30	NDCG@20
Anserini (Lucene)	0.2531	0.3102	0.4240
ATIRE	0.2184	0.3199	0.4211
ielab	0.1826	0.2605	0.3477
Indri	0.2388	0.2995	0.4041
OldDog	0.2434	0.2985	0.4002
Pisa	0.2534	0.3120	0.4221
Terrier	0.2363	0.2977	0.4049

Robertson et al.

Equation 2.1 shows the original formulation of BM25: N is the number of documents in the collection, df_t is the number of documents containing term t, tf_{td} is the term frequency of term t in document d. Document lengths L_d and L_{avg} are the number of tokens in document d and the average number of tokens in a document in the collection, respectively. Finally, k_1 and b are free parameters that can be optimized per collection.

$$\sum_{t \in q} log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right) + tf_{td}} \tag{2.1}$$

Lucene (default)

$$\sum_{t \in q} \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_{d\text{lossy}}}{L_{avg}}\right)\right) + tf_{td}} \tag{2.2}$$

Lucene (accurate)

$$\sum_{t \in q} \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \tag{2.3}$$

ATIRE

$$\sum_{t \in q} \log\left(\frac{\mathbf{N}}{\mathbf{df_t}}\right) \cdot \frac{(\mathbf{k_1}+\mathbf{1}) \cdot tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \tag{2.4}$$

BM25L

$$\sum_{t \in q} \log\left(\frac{\mathbf{N}+\mathbf{1}}{\mathbf{df_t}+\mathbf{0.5}}\right) \cdot \frac{(\mathbf{k_1}+\mathbf{1}) \cdot (c_{td}+\delta)}{k_1 + (c_{td}+\delta)} \tag{2.5}$$

BM25+

$$\sum_{t \in q} \log\left(\frac{\mathbf{N}+\mathbf{1}}{\mathbf{df_t}}\right) \cdot \left(\frac{(\mathbf{k_1}+\mathbf{1}) \cdot tf_{td}}{k_1 \cdot \left((1-b) + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} + \delta\right) \tag{2.6}$$

BM25-adpt

$$\sum_{t \in q} \mathbf{G_q^1} \cdot \frac{(\mathbf{k_1'}+\mathbf{1}) \cdot tf_{td}}{\mathbf{k_1'} \cdot \left((1-b) + \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} \tag{2.7}$$

TF $l \circ \delta \circ p \times$**IDF**

$$\sum_{t \in q} \log\left(\frac{\mathbf{N}+\mathbf{1}}{\mathbf{df_t}}\right) \cdot \left(\mathbf{1} + \log\left(\mathbf{1} + \log\left(\frac{tf_{td}}{1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)} + \delta\right)\right)\right) \tag{2.8}$$

2.4 Results

We found that

2.5 Conclusion

We can conclude that

We [3]

Chapter 3

From Tables to Graphs

Chapter 4

Data Modeling using Graphs

Chapter 5

Applications

Chapter 6

Conclusion

Bibliography

[1] Arguello, J., Crane, M., Diaz, F., Lin, J., and Trotman, A. Report on the sigir 2015 workshop on reproducibility, inexplicability, and generalizability of results (rigor). *SIGIR Forum 49*, 2 (jan 2016), 107–116.

[2] Clancy, R., Ferro, N., Hauff, C., Lin, J., Sakai, T., and Wu, Z. Z. The SIGIR 2019 Open-Source IR Replicability Challenge (OSIRRC 2019). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2019), SIGIR'19, Association for Computing Machinery, p. 1432–1434.

[3] 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.

[4] 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