Survey of Medical Information Retrieval

Prateek Garg
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi, India 110016
prateekgarg.iitd@gmail.com

Shubh Jaroria
Department of Computer Science and Engineering
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi, India 110016
shubh.jaroria@gmail.com

Introduction

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." - Manning, Raghavan, and Schütze, 2008. In a layman's language, it is like looking for needle(s) in a haystack. The need for effective IR techniques grows as the data become more complex and large. The number of people searching for health related information is increasing continuously and it is already one of the most searched topics on the web which makes it an important domain of IR. In this paper, we focus on these medical applications and see where the current state-of-the-art stands.

Electronic Health Record

It is a digital database of people's health records not only limited to one doctor but multiple with an aim of providing improved patient care and health outcomes. It includes their demographics, test results, medical history, history of present illness (HPI), and medications. With increase in EHRs, there has been an increased demand for better information retrieval systems so that doctors need not search a big amount of medical journal data for a small piece of information and thus contributing to more efficient healthcare.

Challenges with Medical IR

A great diversity of users based on the following criterion makes this domain particularly challenging:

- *Varying Information Needs*: This is could be seen by a simple illustration as when a novice researcher/doctor would treating a patient, he/she will need specific facts immediately, then more detailed information when determining a plan of action. On the contrary, an expert researcher/clinician will aim through a complete list of related cases or academic articles related to the patient's disorder or problem. One of the pillars of clinical IR is this level of distinct categories of consumers and their information needs. Perhaps the best measure is the improvement of efficient, theoretically personalised systems that meet these needs.
- *Varying Medical Knowledge*: The users of the Medical IR system could be of various knowledge levels which would be reflected in the way they pose their queries to the system. We need a generalized framework which could adapt itself in order to vary the return solutions as per the complexity of queries asked and user knowledge base. This itself is a difficult challenge.
- *Varying Language Skills*: The research in this domain has been predominantly occupied by English language while there are a large number of medical records and speakers of other languages. This restricts usage of such frameworks by non-English speakers and limits the data required for researching a generalized solution.

While the above challenges could be generalized to other domains of IR as well, the fact that the effect of health on the society is substantial, makes this particular problem especially important and difficult. Further contributing to the difficulty is the lack of proper defined or standardised baselines and small test sets. One of the most recent and popular test sets in the domain is the *NFCorpus* (Boteva et al., 2016) dataset which extracts the data from publicly available sources like *PubMed*.

Data Sources in Medical IR

Information Retrieval focuses on knowledge-based information sources. For medical IR, we can classify the information sources in the following classes:

Preprint. Under review.

- *Primary Sources*: This class of information source consists of original research that appears in journals, books, reports, etc and which deals with initial discovery of health knowledge.
- *Secondary Sources*: These sources are derived from the primary sources and consists of their reviews and condenses. The most common examples of this type of literature are books, monographs, review articles in journals and other publications and opinion based editorials in the form of articles or small blogs.

Recent works

Novel Ways to Retrieve Medical Records

Overview

In [3], an IR system is detailed which is based on reusing the information whether a person finds a document relevant (pseudo-relevance feedback), and using that information to create a user profile, which is used along with a Genetic Algorithm to provide results.

The problem the paper aims to solve is non-uniformity and non-accessibility of medical data for a lot of different hospitals and institutes. Further, this data is stored in multiple different text formats as well, making it hard to create a single unified IR system.The positive user profile basically stores terms and frequencies for documents which the user marked as positive, and stores sensitivity and adjusted sensitivity as a parameter. The negative user profile is similar.

The GA is then used to make use of the top M terms from the positive user profile, and M-N bits from negative user profile, to create a custom query string. This query string, along with the original search query is now used to search for top-K relevant documents to give to the user (which she can mark as relevant/non-relevant, and the user profile is again updated, and the iteration continues).

Results

Although the authors claim that their method obtains an average of 76% relevant documents, there is little to no data supporting their claim. There is no specific knowledge of that dataset they used; there is only a tailored percentage graph which shows the number of relevant documents as a percentage of total documents for each user they tested on.

Further Scope

However, it does seem like an interesting paper with lots of scope for future research - they have demonstrated a method which works, and could potentially be used on any system without worrying about specific data formats. They say that although it doesn't yet take word meanings (ontology) into account, this is an area where work can be done to ensure that it becomes as robust as other models while still being able to work on most data systems accurately.

Tracing the Impact of Hyperparameters on Medical IR Systems

Overview

Even though TREC hosts the Precision Medicine track each year, and teams compete to get higher and higher scores, no one can say with surety exactly why some systems work better than ever. Which features have the highest impact on performance?

In [2], the authors aim to study the impact of hyperparameters on Medical IR systems in a structured way, on two TREC-PM tasks, with two datasets (Biomedical Abstractions, represented by BA, and Clinical Trials, represented by CT). Searching huge spaces of hyperparameters is a big task, but fortunately, previous work in this direction helped to use better methods than a simple grid/line search. After careful consideration, the Sequential Model-based Algorithm Configuration (SMAC) was chosen to tune hyperparameters over other Bayesian methods. Apart from the normal BM-25 parameters, other parameters were also there, for example - different weights for subqueries (disease, gene, age, sex), as well as positive/negative pre-identified keywords. Weights for different diseases as well as their pseudonyms were also treated as parameters. Hyperparameters were found using SMAC, along with ten-fold cross-validation to avoid overfitting. Ablation scores (using only the optimized version of a particular parameter, keeping all others same) were found.

The two different datasets showed different features on which the performance depended, but some commonalities were observed - changing from the phrase query type to the Bag Of Words model led to a sharp drop in nDCG (15.54% for BA and 9.05% for CT).

Results

However, the most important result was that despite using a simple model (BM25 with added features and query expansion), and optimizing the hyperparameters, it wasn't quite far off (1.5 - 2%) from the best reference model submitted.

Future Scope

Such simple, reduced models, which are easier and simpler to maintain and work, can be used as baselines or used for further research, while maintaining upto 98% of the original performance on these test datasets.They show an insight into how important the act of hyperparameter tuning is to even simple models (in this case, BM25F), making them perform in the same league as advanced, state-of-the-art models.

Combining Textual and Image Data for Better Results

Overview

In [1], André Mourão et al aim to enrich the queries for medical records by including images along with textual data to provide better results (both articles and images).

Their approach relies on using expanded queries using Medical Subject Headings (MeSH) thesaurus used with a Simple Knowledge Organization System (SKOS) as input to a BM25L model, along with visual feature extraction (using Local Binary Patterns and High Value Saturation, HSV histograms) from the image query, which is use to create two different rank lists. These rank lists are then combined based on a fusion method based on inverse square rank to provide the final document rankings.

The image retrieval rankings are based on L2 scores between the visual features extracted from the query image and the target images, while the textual rankings are based on the results of the BM25L model after medical query expansion paired with weighted pseduo-relevance feedback.

Results

The model was evaluated on the Visual and Textual Medical Dataset from the ImageCLEFmed 2013 evaluation campaign.

Although this model was able to beat other teams by some margin on the Visual Runs, there was a caveat - the MAP and average precision @10 and @30 was below 5%, indicating that the process is still quite a way off from perfection. The text-only runs were better in this regard, but the runs which used both text and image as inputs were worse off than both, but still better than other teams.

Future Scope

Since this is a novel way of fusion ranking, lots of further research is expected to go in this direction.

Although absolute results are still quite some way off, they are better than other competitors and this gives a lot of hope to maybe use this method one day.

Towards Explainability in IR

Overview

In [4], the authors aim to make IR systems inherently more explainable, and see the results on structured relevance judgement (instead of providing a final ranking based on internal judgement, documents are judged based on intermediate 'aspects' - eg, a decision tree, for a patient who suffers from cancer, is > 50 years old, etc, and use them to provide final relevance ranking) data from the TREC PM track.

Previous methods used query-document pairs to predict ranking scores, reranking documents after an initial retrieval using BM25 (or similar models). An initial retreival stage helps in comparing the performance of the reranking algorithms fairly and objectively. For each topic, both disease and gene terms are used as input to the BM25 model to retrieve the top 500 relevant documents, which can then be reranked further using a decision-tree based approach.

They propose a new retrieval method - train a multi class classifier to get intermediate results, map 'aspects' to outcomes - eg, disease can be exact, more general, more specific, no disease, etc.

Aspects	Outcomes	Classifier Features
Relevance to Cancer Treatment	Human PM Animal PM Keywords Not PM	#Human PM Keywords(n) #Animal PM Keywords(n) #Not PM Keywords(n)
Disease	Exact More General More Specific No disease	#query disease match(n) #disease descendants match(n) #disease ancestors match(n)
Gene	Exact Missing Gene Missing Variant Different Variant	query gene & alisases match(n) is variation in query(b) #query variation match(n) #other variations match(n) is other info in query(b) #other info match(n)
Demographic	Match Excludes Not Discussed	is gender mentioned in article(b) is gender different in article(b) is age mentioned in article(b) difference in age(n)

b: binary valued, #: count of, n: real-valued, PM: precision medicine

The keywords are marked using highest TF-IDF scores in the documents relevant to the outcome.

Aspect	Macro-F1	Accuracy
Relevance to Cancer Treatment	0.45	0.58
Disease	0.46	0.59
Gene	0.41	0.46
Demographic	0.48	0.74

Now, a decision tree is used to learn from the structured relevance data, and is split using pre-identified or learned keywords. Leaves contain the relevance level of the document. There are 2 ways:

- **Deterministic walk (or TREE-HARD)**: Branches are followed by the decision tree according to a strict process - the branch with more than 50% probability is taken to be the correct one and followed further downwards, eventually leading to a single leaf.

- **Probabilistic walk (or TREE-SOFT)**: This method uses the probability percentage as the probability of going down that path, similar to a random walk (but with probabilistic values). Thus, each leaf has some non-zero probability of being reached (the product of all probabilities on the path to the leaf).

Results and Future Scope

Tree soft is reported to be better for inaccurate classifiers, as it will still have non-zero probability to get to the correct output leaf. The authors found that their decision tree outperformed complex Learning-To-Rank models in all aspects. However, take this with a pinch of salt. Only a small improvement (0.01 on P@10 and R-prec) was observed over initial BM25, but there is still lots of scope for future research (particularly because it performs well in reranking modes, and it is inherently explainable, on account of the decision tree).

Conclusion

IR, even medical IR, is a huge field in itself, with different challenges. Different approaches, ranging from an ontology-based approach coupled with a vector-space model were looked at, fleetingly in a paper by Aidarus M. Ibrahim et al[6]. Various aspects to IR - ranging from ML and Image Classification methods combined with textual data[1], as well as genetic algorithms[3] have been observed.

With the rise in the amount of data in this Information Age, there is an increasing need to look at both increasingly complex models as well as simpler ones with better tuning of parameters. Erik Faessler et al[2] conducted such a study which aimed to explore the effect of hyperparameters on models, and compared them to the newer, more complex models. Akin to ML, there is also a growing push for explainability in IR. Jiaming Qu et al[4] tried to come up with an inherently explainable model, that, despite being in nascent stages, could prove to be a base for further studies in this direction.

IR, like many other computer science fields, has some very crucial years ahead of it - there is no dearth of data that is ripe for harvesting.

Acknowledgements

The authors would like to thank Dr. Srikanta Bedathur for providing motivation and continued guidance throughout the project.

References

[1] André Mourão, Flávio Martins, and João Magalhães. 2014. Multimodal medical information retrieval with unsupervised rank fusion. In *Computerized Medical Imaging and Graphics 39(2015) 35-45*.

[2] Erik Faessler, Michel Oleynik, and Udo Hahn. 2020. What Makes a Top Performing Precision Medicine Search Engine? Tracing Main System Features in a Systematic Way. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, Xi'an, China*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397271.3401048

[3] Shihchieh Chou, Weiping Chang Chin-Yi Cheng, Jihn-Chang Jehng et al. 2008. An Information Retrieval System for Medical Records Documents. In *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 2008:1474-7*.
DOI: 10.1109/IEMBS.2008.4649446

[4] Jiaming Qu, Jaime Arguello, Yue Wang. 2020. Towards Explainable Retrieval Models for Precision Medicine Literature Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China*. ACM, New York, NY, USA, 5 pages.
https://doi.org/10.1145/3397271.3401277

[5] Hemant Jain, Huimin Zhao, David P. Klemer, Carmelo Gaudioso. 2005.Semantic Retrieval of Medical Records Related to Patient Symptoms. In *WeB 2005, Las Vegas*

[6] Aidarus M. Ibrahim, Hussein A. Hashi, Abdullalem A. Mohamed et al.2013. Ontology-Driven Information Retrieval for Healthcare Information System: A Case Study. In *International Journal of Network Security Its Applications (IJNSA), Vol.5, No.1, January 2013*.

[7] Catherine Arnott Smith, MA, MILS, MSIS, PhD. Information retrieval in medicine: The electronic medical record as a new domain.

[8] William R. Hersh. Information Retrieval for Healthcare. In *Healthcare Data Analytics*.

[9] Lorraine Goeuriot, Gareth J.F. Jones, Liadh Kelly, Henning Müller, Justin Zobel. Medical Information Retrieval: Introduction to the Special Issue. In *Information Retrieval Journal, January 2016*.

[10] Patrick Kierkegaard, PhD , Rainu Kaushal, MD, MPH , Joshua R. Vest, PhD, MPH. Information Retrieval Pathways for Health Information Exchange in Multiple Care Settings. In *The American Journal of Managed Care, Special Issue: Health Information Technology, Volume 20, Issue SP 17*.