Module: Information Retrieval INM305
Name: Mussa Yousef
Course: MSc Data Science
Email: mussa.yousef.1@city.ac.uk

1. Introduction

Motivation

This Information retrieval (IR) project will be carried out to assess different strategies and techniques to research and obtain information on our research topic. I have gained motivation to research information in this field due to the recent diagnosis of my wife, I am therefore a user and my information need was related to the narrative below. I found some challenging aspects to find specific information related to the topic hence aim to use this as a motivating factor for my research here.

Topic and Evaluation Policy

Title: Multiple Sclerosis

Description: Find information that discusses symptoms and treatment of multiple sclerosis in young women. Thus, investigating the impact on lifestyle, motor skills, and pregnancy.

Narrative: A relevant document will discuss Multiple Sclerosis (MS) symptoms, treatments, and education on MS for young women is relevant. Documents which discuss neural damage from lesions in the brain thus discussing mobility loss are relevant. Documents discussing implications on lifestyle, exercise, and pregnancy are relevant. Documents which discuss the geographic impact on multiple sclerosis are relevant i.e weather and climate.

Understanding the nature of our analysis here our knowledge in the medical and healthcare field cannot answer such a research question adequately hence we utilise Belkin et al (1982) ASK (Anomalous States of Knowledge) model to systematically search for the question above. Considering Ingwersen (1992:p116-117) three types which allow us to classify our research, therefore, tailor our search source. Here we have a research type that is classified as a Muddled topical need, although an extensive direction is provided in the narrative, we must produce an extensive facet analysis, to navigate our search successfully.

2. Facet Analysis

We have decided for this IR task to utilise the PICO predefined framework for our facet analysis. PICO is vastly used in the medical field as the concepts which are used accommodate the medical and healthcare field. Extracting key terms in our description and narrative and tailoring this to the PICO framework, we have the following:

(P)- PROBLEM/POPULATION: - Young Women, Worldwide

(I)- INTERVENTION: - Treatments, Climate, Lifestyle

(C)- COMPARISON: - N/A

(O)- OUTCOME: - Multiple Sclerosis, Brain lesion and Neural Damage

Scrutinising each concept, the initial “P” we have initiated extraction for population, choosing the terms “young women” and “Global”, we justify this by narrowing the gender to women further specifying “young” as stated in the question. Furthermore, the description and narrative doesn’t explicitly mention a geographical location. However, mentioning “…geographic impact on multiple sclerosis” expresses no boundary, this could also be classed in the comparison however we will further analyse the effect of this during our initial searches.

Our “I” expresses any interventions in medical terms this is clear, hence we have placed “Treatments”, “Climate” and “Lifestyle”. These have been classed here to investigate the spectrum of potential investigated solutions for our initiated outcomes, all of which are possible medical remedies which summarise the examples requested to be researched in our description.

We haven’t applied any search terms for “C” as comparisons haven’t been explicitly mentioned in the narrative however, we will revisit the idea of comparing different geographical location and evaluate our initial searches.

Our outcome terms here are “Multiple Sclerosis”, “Brain Lesion” and “Neural Damage” these are the potential outcomes. We have all three terms have a different weighting to them and we will experiment with only using the term Multiple Sclerosis and using all three. Multiple Sclerosis is the definite term we are looking for in our retrieval, however Brain lesion and Neural damage are results or causes of multiple sclerosis as mentioned in the narrative hence we will consider these mutually exclusive and as a union in our initial searches.

3. Search Strategies

We used our initial developed facet as our foundational search, we then aim to go through various iterations developing on familiar search tools such as Google to be able to build a successful framework to test on various other search tools. Considering the nature of the domain is health, our initial search insight should bring to light any key terms which are recurring in our searches and should be included in our final search queries. We aim to build search queries which optimise both the ranking search queries and Boolean search queries, therefore not only making sure our terms are as relevant as possible, however aligning the relationship with key terms in our searches will optimise the reproducibility aspect of our strategy.

Our initial searches for using ranking strategy:

(Young Women, Worldwide)
(Treatments, Climate, Lifestyle)
(Multiple Sclerosis, Brain lesion and Neural Damage)

We begin by assessing our initial ranking search on google to gain early insight on our search vocabulary used. Although results provided relatable documents, we found this didn’t encompass the complete narrative. Therefore, we picked out recurring terms which were relatable and influenced our search and utilised the truncation where necessary for terms which influences our search results too much. For instance, from our initial search the terms “Brain lesion” and “Neural damage” provides restriction to our search, realising the more words we place in our search the more we are limiting the search, our aim here is to optimise reproducibility and not to mirror our results. Given that healthcare is constantly being researched and developed including Multiple Sclerosis, allowing room for freedom in results optimises results hence reproducibility here in the information retrieved tailored to the narrative. We therefore used truncation with both “Neural Damage” and “Brain Lesion” and changed these to “Neural*” and “*Lesion*” respectively. These terms are slightly more open and could be paired with several other medical and health care words which would most certainly remain relative.

Furthermore, the narrative mentions words such as “mobility” and “exercise” which we initially aimed to capture related documents using the term “Lifestyle” however we found more documents mentioned literal activities such as “walking” therefore we replaced this with the general term of “Movement”. For our initial Boolean query, we used ProQuest, as we would be able to search relative published documents and experiment with a range of operators. We

findings allowed us to find key terms which we would need to omit. Using the following simple search:

(*Female OR “Young Women” OR worldwide) AND (“Treatment” OR “Climate” OR Movement) AND (“Multiple Sclerosis” OR *Lesion* OR Neural*)

We found that although our search brought back relatable search some of these had mentioned “Multiple sclerosis” however the document itself was tailored to another medical disorder such as “Iron Deficiency”, “Cancer” or “Stroke” all of which share similar traits to Multiple sclerosis however isn’t tailored to the narrative. We, therefore, subjugated using two techniques, keeping in mind reproducibility, using the (NOT) function on the three specific terms but also making sure our terms results are tailored to our specified title topic we used the operation AB (“Multiple Sclerosis”), makes sure the abstract mentioned multiple sclerosis, hence multiple sclerosis will be a key theme in our retrieved documents.

Ranking

Our ranking search was refined and tested on google, however for this experiment we be diversifying our search on various search platforms such as Bing, Google, YouTube and StartPage. We have identified key terms for our vocabulary investigation and initial Boolean search. We will be keeping our search relatively short only keeping key terms, Our ranking search: (Young “women” treatment movement multiple sclerosis)

Boolean

Our Boolean search was refined and tested on ProQuest, however like our ranking search we will also be searching on various platform, all of those mentioned above excluding YouTube. Trip Database as an additional option tailoring medical documentations. Using ProQuest as a baseline, we have a few unique operators which will help us build a search tailored to our narrative. Furthermore, the advanced search and Command line allows us to examine each layer of results in our facet and analyse the effect of each operation used. First and foremost, we tailored our interface, ProQuest provides a unique advantage to tailor database to query is one we will exploit. We narrow our database down to Healthcare, Pharmaceuticals & Biomedical, Diagnostics, Unassigned, Education databases this took the number of databases being searched on down from 62(default) to 55. To build on our strategy here we used Successive Fractions:

{SET1} = {*Female OR “Young Women” OR Worldwide}
{SET2} ={SET1} AND {“Treatment” OR “Climate” OR Movement}
{SET3} ={SET2} AND {AB (“Multiple Sclerosis”) OR *Lesion* OR Neural*}
{SET4} = *Cancer* OR IRON* OR *Stroke*
{SET5} = {SET3} NOT {SET4}

Interestingly our results are not as expected, as we have set words such as “Young Women” and “Treatment” in quotation these have influenced our search too much. Our results have specified these with the secondary notion of finding Multiple sclerosis. We find that having Multiple sclerosis in our abstract doesn’t mean these documents are generally related to multiple sclerosis hence we will change this to a title search. We make the following changes and investigate our search:

{SET1} = {Female* OR “Young* NEAR/1 Wom?n”}

- We change the truncation from before the word Female to after to allow variations i.e. (Females)
- We changed “Young woman” this restricted our search too much therefore we incorporate the word “Young” with a truncation to open up variations i.e. (Younger) and

also search using proximity of the word Wom?n at most 1 word away therefore we make sure the variation of young relates to Wom?n
- We placed the wild card (?) in the word woman to allow variation of “Woman” and “Women”

{SET2} = {SET1} AND {Treatment* OR Climate* OR Move [*5]

- Here we have taken out the quotation marks for the words Treatment and Climate, these forced results of documents which had these words in. Therefore, we used truncation for both words to allow variation.
- “Movement” didn’t have the impact we anticipated, hence we changed this to MOVE[*5] using truncation and limiting variations to potentially the addition of 5 letters e.g: (Moves), (Movement), (Moving)…etc

{SET3} = {SET2} AND {TI (Multiple Sclerosis) OR *Lesion* OR Neural*}

- We changed the Abstract search to Title search for Multiple sclerosis, making sure MS is mentioned in the title

{SET4} = “Cancer” OR “Iron Deficiency” OR “Stroke”

- We instead of used identical words without variations to make sure these are purposefully omitted.

{SET5} = {SET3} NOT {SET4}

Our aim next was to incorporate the term pregnancy and motherhood into our Boolean search, however our search results above picked these up with using the terms, we got results a huge number of documents which discuss “Female reproductive issues in multiple sclerosis”. Although this Boolean search is tailored to our ProQuest, transferring these over to sources such as Google and Bing bring challenges. Not all the operators available on ProQuest is available on other search tools, however our main Boolean operators such as {AND}, {OR} and {NOT} are available, the {TI} operator is similarly available using the {title:} the other sources, hence we are still able to produce close enough Boolean searches. Furthermore, given our other searching tools provide a spectrum of documents such as: webpage, leaflets and Videos, etc we are able to sacrifice operators as still retrieve related document.

Both our ranking and Boolean searches we had 2 approaches to consider, we aimed to firstly keep our search short and concise using vocabulary which encapsulates the narrative and research question. Secondly, minimising restrictive terms and operators where possible, therefor allowing the search algorithm to suggest documents which tailor to our narrative, without explicitly mentioning specific words mentioned in the narrative. Using these approaches, we were able to make sure our search strategies focused on reproducibility.

4. Evaluation

To evaluate our results, we systematically scrutinised our first 10 results of each searching tool used. As we are taking this experiment at home without ‘Real users’ we will not conduct a laboratory-style evaluation, rather we will consider the quality of our results and focus on our users therefore an operational evaluation will be more tailored. We assess the matching of the requirements set by our narrative and research topic. We have so far established our information need, built our search strategies and now we assess the quality of documents retrieved. Evaluation is subjective to the information need and evaluator hence these are the assessments used when assessing each search system:

- Rate of Repeated documents (RT) – Here our assessment of RT is if the domain in which the data is retrieved is repeated or the document retrieved is repeated from a different domain e.g., Document posted on Facebook/Reddit where the original document has already been retrieved in our search.
- Link Broken (LB) - We opened each document retrieved to validate the link.

- Not retrieved (NT) – here we open each link/document to validate like we have for LB however if the document itself isn't found and we are taken to an alternative site, we classify this as NT.
- Spam – We classified spam documents, as documents which are completely outside of spectrum of information needs. Any sort of marketing, business and false information not from a trust worth source is classified as spam.

Subsequently, once we have checked each document for the above requirements against our narrative. We then made the classification on the relevance (Recall) of each document; we classify documents as relevant if they do not fall into the criteria's above and the document relates to multiple sclerosis and has at least one other keyword in the document. Sequentially, we collated for each search system, the precision relevance percentage for top 5 documents retrieved, top 10 documents retrieved and collected the absolute average precision of all retrieved documents. Results obtained are presented in the table below [Table 1 - Below].

Search System	P@5	P@10	AveP	RT	LB	NT	SPAM
Dialog Boolean	0.80	0.90	0.76	0	0	0	0
TripDatabase Boolean	0.80	0.80	0.68	0	0	0	0
Bing Ranking	1.00	0.90	0.90	1	0	0	0
Bing Boolean (2DSearch)	0.80	0.90	0.82	1	0	0	1
Google Ranking	1.00	1.00	1.00	2	0	0	0
Google Boolean (2Dsearch	1.00	1.00	1.00	1	0	0	0
Google Images Ranking	1	0.90	0.88	0	0	0	0
Bing Images Ranking	1.00	0.80	0.80	0	0	0	1
Youtube Ranking	0.80	0.90	0.84	0	0	0	0
Bing Video Ranking	1.00	0.90	0.90	0.00	0.00	0.00	0.00
Social Searcher Ranking	1.00	0.80	0.76	0.00	0.00	0.00	0.00
Startpage Boolean	0.80	0.90	0.82	0.00	0.00	0.00	0.00
Startpage Ranking	0.80	0.50	0.46	0.00	0.00	0.00	0.00

Table 1 – Consolidated Search Results table format

Bing – We have used three Bing platforms to evaluate our ranking strategy, Bing query, Bing Images and Bing Video. All three provided interesting and varying result however it seems that our strategy here was effective with Bing's algorithms. We had success without P@5 with all three providing 100% this suggests that our ranking strategy was tailored to optimise our search, further reinforcing reproducibility. P@5 here for all three passed all four assessments above hence underpinning quality in our top 5 results. Our precision at 10 result although we had 90%,80%,0.9% respectively, we had one repeated document for Bing Query, which suggest possible limitation on information on our search. Bing images presented 1 spam in our P@10 which was related to women but not multiple sclerosis, in this case the term "Woman" may have had too much of an influence on our search. Bing Video, only irrelevant result was from 'man' with multiple sclerosis, all our other terms were highlighted in the result. However, as women is the population, we are searching we didn't consider this as relevant from the user's perspective.

Google- We scrutinised three google platforms to evaluate our ranking search, Google Query, Google images and Youtube. Our Google query search provided success all round with 100% precision in all evaluated indexes. We retrieved documents from the NHS, Healthline, Mayo Clinic and healthy women, all of which are trusted sites and had tailored information on women with MS. However, we did find 2 documents repeated which were both found in our first 5 results, this document was of the same article published from NCBi. These documents are 3rd and 4th ranked respectively, underneath our retrieved link we can see different words are being highlighted, which show which key words influenced the ranking of this retrieved documents, we find different key words highlighted however consistently both have highlighted multiple sclerosis, this could be due to the proximity of the key words at different areas of the article regardless this reinforced the relevance of this document, rather than highlighting the fault of our search strategy. Google images provided interesting diagrams of different types of MS and Geospatial mappings of where high cases of MS are located however our one document which wasn't relevant was of article on movement and physiotherapy, although relevant to a

potential MS patient, this isn't relevant to the potential user here. Youtube also showed great precision however like our Bing multimedia platform we searched on we found two documents relating to men, which doesn't relate to our population.

Noticeably, our social search retrieved results which was of the same links as our google search apart from our 9th and 10th documents which were unrelated videos. Links were direct to the sites which are related to multiple sclerosis. Investing further we find all documents after the 8th document is not related which suggest that multiple sclerosis awareness on social media maybe minimal or not a relatively method to post information. Furthermore, we investigated Startpage a private search source and found similar results as social searcher also rooted to googles retrieved documents. Our first 4 documents are the exact same. This shows the influence googles algorithm has on other platforms. Startpage however retrieved far more irrelevant adverts which related to pharmaceuticals and therapy hence our precision here was adversely affected at 46%.

We utilised the 2Dsearch to carry out our Boolean searches for Bing and Google. Google once again brought back 100% precision for our Boolean search, however there were more documents here related to women and pregnancy compared to our ranking results which is fantastic. We were initially going to include [Preg*] in our Boolean search however our aim was to allow the algorithm respectively to read "in-between the lines" allowing room for flexibility which was successful here. Our Bing search here slightly provided less precision on our top 5 retrieved documents, we found more academic papers here on our Bing search compared to google. Both Google and Bing Boolean searches had the exact same repeated link which was from a relevant NHS webpage directory.

Our ProQuest provided a precision at 5 at 80% one of the documents was of a paper discussing neurological lesions only partly focusing on MS. Our precision at 10 provided a 90% success rate. All our other documents were related to our Facets and all were research papers from PhD or master's students. Our "Other" search source was TripDatabase tailored to healthcare professional, we expected a high precision however we found the research papers found on this document was extremely sophisticated and some of which scrutinise facets of Multiple sclerosis which does not related to the narrative hence our P@10 was 80%.

5. Summary

Overviewing our entire process, we found the Boolean search provided successful search results, however, considering the user's flexibility, the nature of the topic, and rate of research completion. The ranking search method provided excellent results which maximises reproducibility. Our process used for this research was influenced by Ellis (1989) [Diagram 1 – Appendix].

This process allowed us to consider the behaviour of our users using a framework that considers information-seeking behaviours. Throughout this process we tailored each step through cognitive, affective, and situational factors, making sure our search strategies would withstand the element of time hence reproducibility is maximised. We have explored the facets integrated in the information need itself and have learnt the methods in which successful strategies work to optimise results using Ranking and Boolean strategies. Practically our results were reasonable, however optimising our strategies we were able to utilise search engines to work in our favour by implementing key terms and operations. If we could improve this experiment, we would have potentially used a VPN and log into other servers around the globe to explore and compare the different information retrieved from different countries. The population for this experiment was women worldwide hence, although our results in this experiment was successful, our results were tailored to the UK i.e., NHS and MSsocietyUK. How would our retrieval be for the Netherlands, France, etc. This would have been interesting to explore hence this is how we could have improved our research in this area.

Appendix:

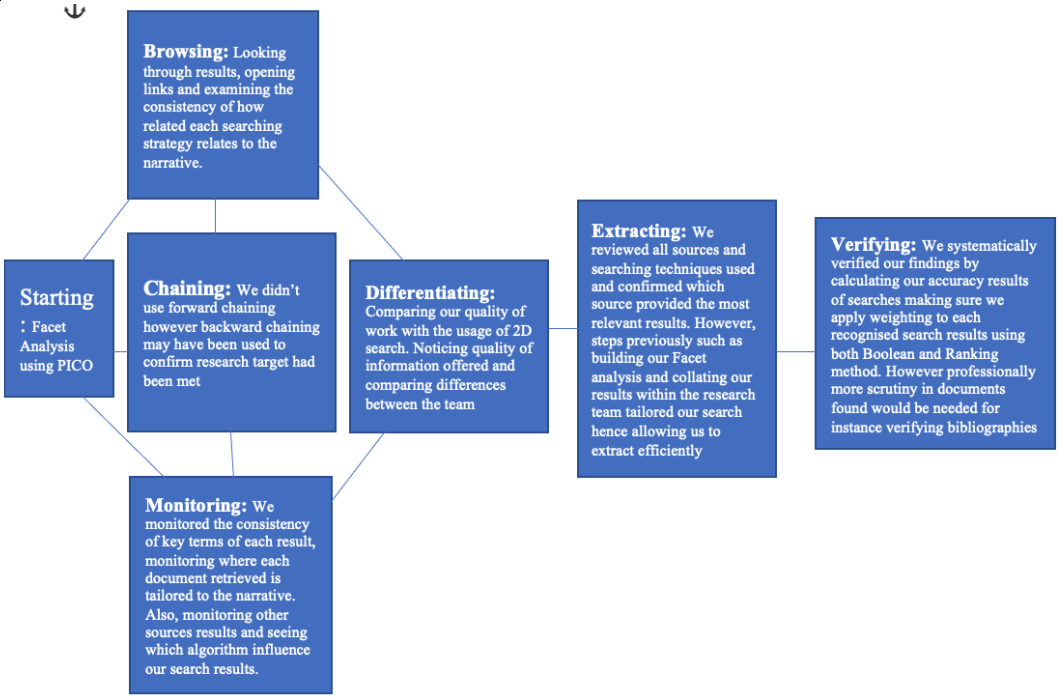

Diagram 1 – Ellis evaluation model

References: (No particular order as we didn't explicitly reference a paper in this report apart from the ones mentioned in the report)

[1] Boudin F., Shi L., Nie JY. (2010) Improving Medical Information Retrieval with PICO Element Detection. In: Gurrin C. et al. (eds) Advances in Information Retrieval. ECIR 2010. Lecture Notes in Computer Science, vol 5993. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12275-0_8

[2] Cooper, C., Booth, A., Varley-Campbell, J. *et al.* Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. *BMC Med Res Methodol* **18,** 85 (2018). https://doi.org/10.1186/s12874-018-0545-3

[3] Dr Andrew Macfarlane – Information Retrieval INM305 Lecture Notes.
Lecture 1 – Lecture 10