



Meta Search Engine using Semantic Similarity and Correlation Coefficient

Naresh Kumar, Deepak Sharma, Nripendra Narayan Das

Abstract- This paper aims to provide an intelligent way to query and rank the results of a Meta Search Engine. A Meta Search Engine takes input from the user and produces results which are gathered from other search engines. The main advantage of a Meta Search Engine over methodical search engine is its ability to extend the search space and allows more resources for the user. The semantic intelligent queries will be fetching the results from different search engines and the responses will be fed into our ranking algorithm. Ranking of the search results is the other important aspect of Meta search engines. When a user searches a query, there are number of results retrieved from different search engines, but only several results are relevant to user's interest and others are not much relevant. Hence, it is important to rank results according to the relevancy with user query. The proposed paper uses intelligent query and ranking algorithms in order to provide intelligent meta search engine with semantic understanding.

Keywords: Meta Search Engine, Scrapping unit, Intelligent Query System, Rankings

I. INTRODUCTION

A human can see and understand words and sentences. But to make the words and sentences understandable by machines, these have to represent them in form of vectors. A machine will be able to get similar words based on particular word by representing the words as vectors. A machine will be able to do so by using some similarity metrics that can be cosine similarity or dot similarity. There are various pre-trained models that can be used to convert word into vectors like word2vec and glove2vec.

These are the models which can convert word to vectors. But these models can't give us the encodings for the sentences directly. Author find encoding for the sentences indirectly by finding out the vector of each word in the sentence using either of these models using word2vec or glove2vec and averaging them. This will give us the vector which will show the meaning of sentence in a vector form.

But here is one major problem and that is while finding the vector form of sentence author have not considered the word semantic order inside the sentence. This semantic word order is very important in consideration for conversion into vector form of sentences.

For example there are 3 sentences:

1. How old are you?
2. What is your age?
3. How are you?

It is obvious that 1 and 2 are semantically the same even though 1 and 3 have more common words. A good sentence encoder will encode the three sentences in such a way that the vector for 1 and 2 are closer to each other than 1 and 3. So to eliminate this problem Universal Sentence Encoder by Google which is a very powerful model for encoding the sentences which will preserve it's semantic word order.

Universal Sentence Encoder is able to embed not only words but also phrases and sentences. It takes variable length English text as input and outputs a 512-dimensional fixed vector.

II. LITERATURE SURVEY

Authors of [1] consider the architectures and features of Meta Search Engines (MSE) for extracting documents from one or more domains on the web. It analyzes two MSE i.e. general MSE and Special purpose MSE. Authors of [2] proposed a MSE on the basis of clustering and ranking to find the relevant results. It have user interface, relevancy calculator, cluster generator and webpage adjuster as a main modules. It took top 10 results Search Engine (SE) –Bing, Google, and Alta Vista) and tested the proposed work on 30 different queries. Traditionally, MSE matches the query with the webpage and then provide the result which is as keyword matching. According to [3], a MSE uses interfaces of self forward query to SEs produce its results from the Internet. MSE consume input form end user and simultaneously send out this query to multiple SEs for getting results. Then it formats the received result and present it to end user. Set of keywords in a single query and Word Net ontology is used to provide the most suitable query to SE through the MSE is proposed in [4]. Some authors used optimization techniques [5], some are using query semantics and some are using, clustering techniques [6] to improve the architectures of MSEs but still they have challenges [7]

III. ISSUES AND CHALLENGES

Problem in [4] is that the author is making the query set by extracting the similar words to the given query using Word-net Ontology. Problems is that:

- A. **Large number of permutations with synonyms:** Taking the words from the sentence, finding the synonyms of words and making a query set using synonyms of words. The number of sentences formed by using these synonyms will be very high.

Revised Manuscript Received on July 30, 2020.

* Correspondence Author

Naresh Kumar*, Associate Professor, Department of CSE, Maharaja Surajmal Institute of Technology, New Delhi, India. E-mail: narsumsaini@gmail.com

Deepak Sharma, Assistant Professor, Department of IT, Jagannath International Management School, Vasant Kunj, New Delhi, India, E-mail: deepaktech@hotmail.com

Nripendra Narayan Das, Associate Professor, Department of Information Technology, Manipal University, Jaipur, India. E-mail: nripendradas@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

We will do this for all the result pages and sort pages according to the correlation distance factor. After sorting results will be shown to the users.

Dataset used & Correlation Distance Factor: This paper used quora questions pair’s dataset for finding out the similar queries. This dataset can be taken from the kaggle website. This dataset was published on 2016. There are over 400,000 lines of potential question duplicate pairs. This makes up total of 800,000 questions. After this each question is input to our universal sentence encoder model to get the corresponding 512 dimensional encoding vector which is stored in the database. It is a correlation coefficient between 2 vectors which will tell us how much these vectors are similar to each other. This factor will give us a value in between (0-1). If this factor is approaching to 1 then that means 2 vector are very similar to each other, otherwise they are not. Usually we can have to take a threshold value for this factor which can be 0.9 or 0.8 to distinguish between similar and not similar vectors. Here, in our implementation we have taken the threshold value to be 0.8. Universal Sentence Encoder will give us the encoding of vectors which is a fixed dimensional vector of size 512 irrespective of the size of input, that means whether the word, sentence or even paragraph is fed into the encoder it will give us a fixed length vector which will explain the input (of any length) in 512 dimensional space. With the help of this we can easily find the relevancy between the queries and between the query and pages. Output of Encoder which is 512 fixed dimensional vector will help us to main word semantic order in sentence.

Results and Discussion: We have taken 5 queries from 5 domains (Person, Music, Education, Social Media, General). Queries that we have taken are:

- (i) Who is Barack Obama? (Person)
- (ii) Which are the best rap songs? (Music)
- (iii) What is Web Applications? (Education)

How can we use twitter for money? (Social Media) Is creativity important? (General) Results for each query are shown in the table below. Table contains 4 columns (query, similar queries, results, relevancy score for each result). Each column is described below in brief:

- (i) Query: this is the query input to the MSE.
- (ii) Similar Queries: Similar queries of a given Query from database.
- (iii) Results: Final extracted resultant pages
- (iv) Relevancy Score: Score that represents correlation between original query and contents of resultant links.

Table 1 shows the following:

- (i) Shows the results for the query “Who is Barack Obama?”
- (ii) Shows the results for the query “Which are the best rap songs?”
- (iii) Shows the results for the query “What is Web Applications?”
- (iv) Shows the results for the query “How can we use twitter for money?”
- (v) Shows the results for the query “Is creativity important?”

| S. No. | Query | Similar Queries | Result | Relevancy Score |
|--------|----------------------|--------------------------------------|---|-----------------|
| 1 | Who is Barack Obama? | Who is the president of America? | https://www.biography.com/people/barack-obama-12782369 | 0.41393143 |
| | | Who is the President of America now? | https://www.history.com/topics/us-presidents/barack-obama | 0.41218776 |
| | | | https://barackobama.com/ | 0.3825521 |

| | | | | |
|---|-------------------------------|---|---|------------|
| | | | https://en.wikipedia.org/wiki/Barack_Obama | 0.28192896 |
| | | | https://en.wikipedia.org/wiki/President_of_the_United_States | 0.25138646 |
| | | | https://en.wikipedia.org/wiki/Bill_Clinton | 0.23507671 |
| | | | https://en.wikipedia.org/wiki/Presidency_of_Donald_Trump | 0.22485167 |
| | | | https://www.britannica.com/biography/Barack-Obama | 0.20174456 |
| 2 | Which are the best rap songs? | What are some good rap songs to dance to? | https://www.thetoptens.com/rap-songs/ | 0.69754565 |
| | | What are some of the best rap songs? | https://digitaldreamdoor.com/pages/best_rap-songs.html | 0.5860768 |
| | | | https://www.funadvice.com/q/dance_148283 | 0.5260219 |
| | | | https://www.thoughtco.com/best-rap-songs-of-all-time-2857834 | 0.49227604 |
| | | | https://www.thoughtco.com/top-rap-songs-of-the-90s-2858039 | 0.46373957 |
| | | | https://www.thoughtco.com/best-hip-hop-dance-songs-2858028 | 0.46275324 |
| | | | http://www.madmanmike.com/dance_songs.html | 0.43373996 |



| | | | | |
|---|-----------------------------------|--|---|--------------------|
| | | | http://colemizestudios.com/blueprint-for-writing-rap-songs/ | 0.425 2886 2 |
| 3 | What is Web Applications? | What is web application? | https://rubyonrails.org/ | 0.558 0404 |
| | | What is the web application framework? | https://en.wikipedia.org/wiki/Web_application | 0.506 0072 |
| | | | https://en.wikipedia.org/wiki/Web_framework | 0.442 0372 5 |
| | | | https://docs.microsoft.com/en-us/aspnet/whitepapers/add-mobile-pages-to-your-aspnet-web-forms-mvc-application | 0.441 9880 8 |
| | | | https://en.wikipedia.org/wiki/Web_application_framework | 0.440 8918 |
| | | | https://searchsoftwarequality.techtarget.com/definition/Web-application-Web-app | 0.438 9917 6 |
| | | | https://www.geeksforgeeks.org/top-10-frameworks-for-web-applications/ | 0.427 1093 3 |
| | | | https://searchsoftwarequality.techtarget.com/.../Web-application-Web-app | 0.408 2357 |
| 4 | How can we use twitter for money? | How do I use Twitter as a business source? | https://www.jeffbullas.com/37-ways-to-use-twitter-for-business/ | 0.480 7621 5 |
| | | How can I use Twitter for business? | https://www.socialmediaexaminer.com/how-to-use-twitter-for-business-and-marketing/ | 0.475 8970 7 |
| | | | https://www.businessnewsdaily.com/7488-twitter-for-business.html | 0.465 7090 3 |
| | | | https://www.lifehack.org/articles/money/7-creative-and-effective-ways-make-money-twitter.html | 0.388 5359 5 |
| | | | https://analytics.twitter.com/ | 0.366 0281 3 |
| | | | https://www.youtube.com/watch?v=t9Si-UEczvM | 0.362 2422 2 |
| | | | https://www.wikihow.com/Earn-Money-Using-Twitter | 0.334 4718 8 |
| | | | https://business.twitter.com/en/basics/intro-twitter-for-business.html | 0.325 5081 2 |

| | | | | |
|---|--------------------------|------------------------------|---|--------------------|
| 5 | Is creativity important? | Is Creativity Important? | https://www.elitedaily.com/money/entrepreneurship/creativity-important-quality | 0.591 7218 |
| | | Why is creativity important? | https://www.linkedin.com/pulse/importance-creativity-innovation-business-siyana-sokolova | 0.399 2474 7 |
| | | | https://alistemarketing.com/blog/reasons-why-creativity-is-important-to-decision-making/ | 0.290 2453 2 |
| | | | https://en.wikipedia.org/wiki/Creativity | 0.289 6288 6 |
| | | | http://www.edudemic.com/creativity-in-the-classroom/ | 0.258 1226 2 |
| | | | http://www.innovationmanagement.se/imt-ool-articles/why-diversity-is-the-mother-of-creativity/ | 0.245 0365 1 |
| | | | https://tscpl.org/art/why-is-creativity-important-in-everyday-life | 0.203 5246 5 |
| | | | https://www.dictionary.com/browse/creativity | 0.111 4332 6 |

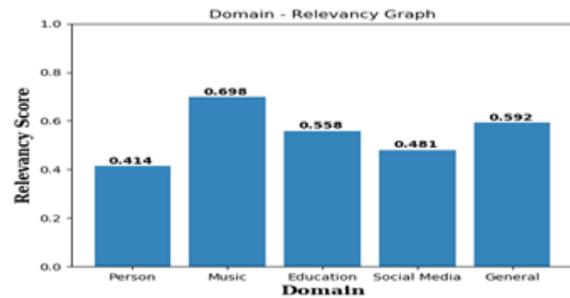


Fig. 2. Domain vs. Relevancy scores

VII. CONCLUSION

In proposed MSE authors have minimized the issues of intelligent query system by using Universal Sentence Encoder instead of Word2Vec model. This will help us to get similar semantic sentences. This helps us to maintain word semantic order in the sentence. In this MSE extraction of similar queries to the original query using correlation factor and then get the results from 3 search engines (Google, Bing, Yahoo) for each query. After getting results, ranking is performed according to the relevancy score. Relevancy score is calculated using correlation between contents of pages and the original query.

REFERENCES

1. A. Madhavi and K. Harisha Chari, "Architecture Based Study of Search Engine and Meta Search Engines For Information Retrieval", International Journal of Engineering Research & Technology (IJERT), Vo.2, Issue 5, ISSN: 2278-0181, May – 2013
2. N. Kumar and R. Nath, "A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering," International Journal of Computer, vol. 10, issue. 1, pp.1-7, ISSN: 2307-4531, 2013.



3. The Wikipedia website [Online]. Available: http://en.wikipedia.org/wiki/Meta-search_engine at 4:00PM (IST) on 23/09/2018.
4. V. Sivakumar, "Semantic Meta Search Engine Using Semantic Similarity Measure", International Journal of Information System and Engineering, Vol. 3, ISSN: 2289-7615, November, 2015
5. Mr. K. P.Raghuvashi, "An Empirical Study on the Meta- Search Engine Optimization Technique Based on Keyword: A Review", IBMRD's Journal of Management and Research, Vol. 3, Issue-2, ISSN: 2277-7830, September 2014
6. N. Kumar, Sonali and S. Gupta, "Non Overlapping Clustering based Meta Search Engine", International Journal on Future Revolution in Computer Science & Communication Engineering, ISSN: 2454-4248, Volume: 3 Issue: 8, 172 – 177.
7. The Bing website [Online]. Available: https://blogs.bing.com/search-quality_insights_May-2018/Towards-More-Intelligent-Search-Deep-Learning-for-Query-Semantics at 4:00PM (IST) on 23/09/2018.

AUTHORS PROFILE



Naresh Kumar holds a Ph. D. from Kurukshetra University, Kurukshetra and M. Tech. (Computer Science & Engineering) degree from YMCA University of Science and Technology, Faridabad. He is currently working as an associate professor at Maharaja Surajmal Institute of Technology, New Delhi since 2011. His area of research interest

includes web crawlers, search engines and meta search engines. He has published more than 44 research papers in reputed journal and conferences.



Deepak Sharma is currently working as an Assistant Professor in the Department of Information Technology, Jagannath International Management School, Vasant Kunj, New Delhi. I am pursuing his PhD(CS) from Jagannath University, Jaipur. Also obtained M.Tech(CS) and MCA degree. I am having

more than 14 years of experience in academics and industry. Main interest areas include Programming, Data Structure, Data Mining and Machine Learning, Search Engines and Web Crawlers. During the tenure of my service, I hold various important positions and proved myself as valuable asset for the organization. I was awarded "Best Employee Award" by the institute & also received appreciation letter from the Education department of Kandahar province, Afghanistan.



Nripendra Narayan Das received PhD from Gautam Buddh University, UP, India. Currently He is working as an Associate Professor in Department of Information Technology, Manipal University Jaipur, India. He has published more than 30 papers in National and International Journals. He has more than 24 Years of experience in Industry as well as in teaching.