Goslin, Kyle and Hofmann, Markus (2019) A Qualitative Analysis of the Wikipedia N-Substate Algorithm's Enhancement Terms. Journal of Computer-Assisted Linguistic Research 3: 67-77.

A Qualitative Analysis of the Wikipedia N-Substate Algorithm's Enhancement Terms

Kyle Goslin*, Markus Hofmann
Department of Informatics and Engineering Technological University Dublin, Ireland.
*Corresponding author: Kyle.Goslin@itb.ie

Received: 23 December 2018 / Accepted: 26 February 2019 / Published: 15 July 2019

Abstract
Automatic Search Query Enhancement (ASQE) is the process of modifying a user submitted search query and identifying terms that can be added or removed to enhance the relevance of documents retrieved from a search engine. ASQE differs from other enhancement approaches as no human interaction is required. ASQE algorithms typically rely on a source of a priori knowledge to aid the process of identifying relevant enhancement terms. This paper describes the results of a qualitative analysis of the enhancement terms generated by the Wikipedia N-Substate Algorithm (WNSSA) for ASQE. The WNSSA utilises Wikipedia as the sole source of a priori knowledge during the query enhancement process. As each Wikipedia article typically represents a single topic, during the enhancement process of the WNSSA, a mapping is performed between the user's original search query and Wikipedia articles relevant to the query. If this mapping is performed correctly, a collection of potentially relevant terms and acronyms are accessible for ASQE. This paper reviews the results of a qualitative analysis process performed for the individual enhancement term generated for each of the 50 test topics from the TREC-9 Web Topic collection. The contributions of this paper include: (a) a qualitative analysis of generated WNSSA search query enhancement terms and (b) an analysis of the concepts represented in the TREC-9 Web Topics, detailing interpretation issues during query-to-Wikipedia article mapping performed by the WNSSA.

Keywords: Automatic Search Query Enhancement, Text Analysis, Wikipedia.

1. Introduction

When a user is interacting with a search engine, typically they enter terms that they deem to be relevant to the content they want to retrieve. These search queries often do not contain any consistency in formatting, length, domain or spelling. In addition to this, users often have difficulty expressing what they are trying to find in search terms. An early analysis of search queries by Jansen et al. (1998), showed an average of length of 2.35 terms, with 80% of the

DOI: https://doi.org/10.4995/jclr.2019.11159
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA. Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

identified queries consisting of three or less tokens long. A more recent study by Mastora et al. (2008) showed 57.7% of analysed queries being only one term long. This shows a trend towards the dependence users are placing on the ASQE algorithm to aid the search process.

This creates a difficulty during the development of ASQE algorithms as a 'one size fits all' approach must be used for developed algorithms, avoiding overfitting to a particular domain or user profile. As a solution to these problems, an algorithm was developed titled the Wikipedia N Sub-state algorithm (Goslin, 2017) that uses up to date Wikipedia content as a source of a priori knowledge for candidate term collections and term weighting.

In this paper, Section 2 provides a background to ASQE and the algorithm results under analysis, the WNSSA. Section 3 provides a qualitative analysis of the enhancement terms that were generated by the WNSSA with a focus on the queries that performed poorly during relevance assessment. This is followed by a review of the concepts/articles that were represented in the first iteration of the algorithm. Based on these results, Section 4 provides a brief discussion and analysis of the findings. Section 5 concludes this research reviewing results gathered from this analysis, posing questions for future research.

2. Background

ASQE algorithms are designed to automatically enhance a query to make search results from an information retrieval engine more relevant. This process is completed irrespective of the experience the user has. However, the length of a search query entered by a user is often the only piece of information available to aid the ASQE algorithms. In a study by Bazzanella et al. (2010), 4017 queries were analysed finding an average length of 2.04 terms. In these queries, over 35% contained one term and less than 3% contained five or more terms. From this limited size, context of queries often becomes an issue for ASQE. A number of methods have been used to add context during ASQE including user profile information (Asfari et al. 2009), query log data (Gao et al. 2013) and thesauri (Voorhees, 1994). In addition to this, an issue can be seen with the number of enhancement terms generated by an ASQE algorithm. Ogilvie et al. (2009) outlined that ten or fewer terms provided the best Average Precision, and that there is no single perfect number of enhancement terms for all queries.

The proposed algorithm, the Wikipedia N Sub-state Algorithm (WNSSA) was designed to add context and generate suitable enhancement terms for any given user search query. This algorithm contained two core variables, states and sub-states. The states variable represents the overall number of iterations the algorithm performs on a given query *Q*. For each iteration of the algorithm, N number of sub-states are run internally by the algorithm. During these sub-states, additional terms can be added into the sub-process, further broadening the collection of terms before it is passed to the next state before final enhancement terms are generated. The iterative nature of the WNSSA allows a search query or stem term to be taken and then used to find terms that are relevant in Wikipedia. By doing this the terms that would not have previously been identified can be connected to the stem through their association with a located related Wikipedia article.

The WNSSA solely relies on Wikipedia for this knowledge. One of the main advantages of

using Wikipedia as a sole source of a priori knowledge is the rich collections of terms that would typically not be found in other sources. In addition to this, descriptive URLs are often created for common titles and phrases. To further this, a large collection of redirects is present in Wikipedia redirecting one article to another article even if spelling errors or incorrect prefixes are used. This provides an opportunity to make a mapping between search queries entered by users and the title of Wikipedia articles directly. When this fails, the WNSSA proceeds to tokenize the query as a last attempt to make a mapping between query terms and possible Wikipedia article URLs. The WNSSA also utilises the backlink API and search API available to aid enhancement.

As the WNSSA allows for parameterisation, Table 1 describes each of the chosen parameters for the analysed run which have shown a high performance during previous analysis. During the algorithm run, 10 overall iterations are performed with 5 internal sub-states. The approach used by this algorithm was set to append term, informing the algorithm not to replace the original terms but to add on to the existing query *Q*. To replicate the real-world search, the TREC-9 Web Topic collection was used on the ClueWeb12 data set. The TREC-9 data set contains 50 search topics that contain spelling, domain and length variation to replicate real search queries.

Parameter	**Value**
States	10
Sub-states	5
Terms per sub-state	1
SearchLax	0
Approach	Append Term
Term Window Size	2

TABLE 1. WNSSA PARAMETER SETTTINGS FOR EVALUATED RUN.

For each enhanced search query, the query was submitted to the ClueWeb12 data set. This data set consists of 733,019,372 English web pages, collected between February 10, 2012 and May 10, 2012. After a query is submitted, 100 search results are returned. For each of the top 10 results that were returned, the precision was calculated, shown in Equation 1.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

EQUATION 1. CALCULATING PRECISION FOR QUERIES.

Once the precision can be calculated, the Average Precision (AP) can then be calculated. This is the average of the ten precision scores for query *Q*, which also considers the ranking of each of the documents created during the retrieval process. In the following equation, *k* is the current rank in the sequence of retrieved documents. *n* is the number of retrieved documents and *P(k)* is the precision for the document at position *k* in the list. *rel (k)* is the relevance score which is either 0 if not relevant to the original query or 1 if relevant.

$$AP = \frac{\sum_{k=1}^{n}(P(k)\ X\ rel\ (k))}{number\ of\ relevant\ documents}$$

EQUATION 2. CALCULATING AVERAGE PRECISION (AP).

3. Analysis

Based on the configuration described in Section 2, Section 3.1 provides a look at the generated enhancement term relevance, identifying which terms added to the original query were relevant or not and Section 3.2 which details the concepts identified during the enhancement process for each query by the WNSSA. In the collection of enhancement terms generated by the WNSSA, many conceptually related terms can be seen. There are often cases, however, when terms are added that appeared frequently in the collection of stem articles (articles related to the query that were used during the enhancement process) but did not have a direct impact on the precision score for the enhanced query. FIGURE 1 and FIGURE 2 outline the average precision (AP) scores for each test topic enhanced with the terms generated by the WNSSA. Although some topics may have had a high AP score, it is often the case that poor-quality terms may have been part of the enhanced query. From this, the overall AP score can be reduced.

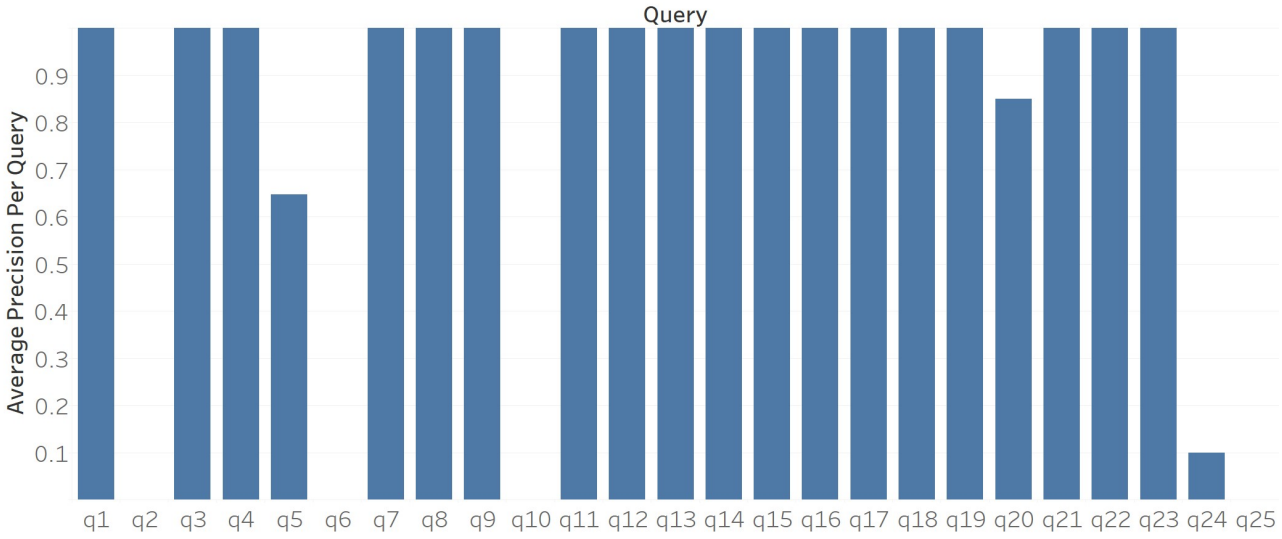

FIGURE 1. WNSSA AVERAGE PRECISION SCORES FOR QUERIES 1-25.

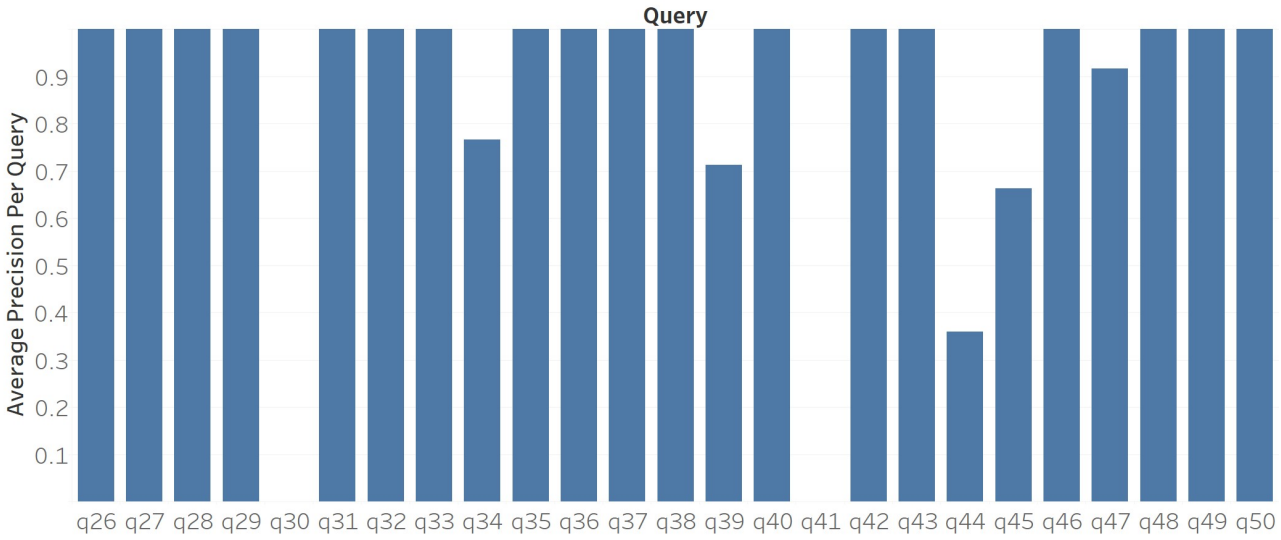

FIGURE 2. WNSSA AVERAGE PRECISION SCORES FOR QUERIES 26-50.

3.1 Generated Enhancement Term Relevance

The generated enhancement terms are shown in TABLE 1. In query 2 ('do beavers live in salt water'), although many of the terms that were generated were highly relevant to the area of beavers and the location that they live, additional terms were added that impacted the overall precision. The terms 'state, north, area' and 'city' caused issues during the retrieval process. Query 6 ('is the world going to end 2000?') caused a number of issues. In the generated results we can see that as no direct mapping was made to a Wikipedia article. An appropriate mapping would be to /wiki/ Year_2000_problem that identifies the issues that would appear in the year 2000. Many of the terms in this original search query were treated as stop words. As this mapping was made to the article /wiki/2000. This caused a number of high ranking irrelevant terms to be utilised.

Query 10 ('Who was Moses?') posed similar issues with stop words. In the first half of this query, 'who' and 'was' were treated as stop words leading 'Moses' to be used as the primary stem word for article mapping. Typically, those with a religious background would consider Moses to be the man referenced in biblical times. When additional iterations of the WNSSA were placed on the single term 'Moses', the term was mapped to the Wikipedia article /wiki/Moses. In this article there are multiple references to the city and location where he was believed to be born. References to 'Robert' can be traced to the drawing of Moses by Robert Walter Weir.

Query 24 ('how e-mail bennefits businesses'), originally contained a spelling error that was fixed by the WNSSA during the initial query run. This query contained only one stop word, leaving three core terms for analysis. Again, no direct mapping could be made for 'email', 'benefits' and 'business' so individually these terms were mapped. In the mapped article /wiki/Email, many references to business and benefits could be found. The terms that were gathered for the enhancement process including 'marketing', 'messaging' and 'enterprise' are all relevant to the core query. As the question was taken out of context, the results returned from the information retrieval process were poor. Query 25 ('what is the composition of

zirconium?') was another question based query. The most direct mapping for this query is /wiki/Zirconium. This page, however, had no direct reference to zirconium. Many of the enhancement terms generated are related to zirconium including 'alloys' 'metal' 'compositions' and the environments where it is used including 'nuclear'. These enhancement terms can be considered high quality, but the question-based nature of the query provided poor results during IR. Query 30 ('car traffic report') did not consist of any stop words. When the enhancement terms are relevant excluding 'original' 'radio' and 'indicate'. Although the majority of these terms were relevant, the context of the query was lost. Rather than reading a collection of car traffic reports, the disambiguation of the terms caused the query to fail. Even if the ideal Wikipedia was mapped, e.g., /wiki/Traffic_reporting was reached, the context of returning traffic reports was lost.

ID	TREC-9 Web Topics with Enhancement Terms
1	bengal cat breed leopard domestic
2	beaver live salt water creek park dam **state north area city** lake island **near**
3	hunger **force part** world states **late** food **american based united** organization
4	parkinsons disease foundation target potential treat drugs appears **years** people **model numbers**
5	jackie robinson appear first game league appearance run cincinnati home **years** player games appeared **modern**
6	world end 2000 war **italy following series without new championship president million** country
7	chevrolet trucks colorado **share** suv silverado **task force full similar** design **index**
8	fasting month fasts lent period prayer church food orthodox holy **year**
9	lender foreclose property upon lenders bank borrower repossess money mortgage needs debt home
10	moses **born robert city known states united** hebrew american york **named**
11	lava lamps uk liquid motion random made **howstuffworks** used water tetrachloride **craven**
12	real estate new jersey nj camden profile york neighborhood apartments homes states community united
13	tartan clan register scottish official green scotland red world blue white
14	nativity scene saint jesus christmas culture innocents joseph bethlehem star mary **nicholas**
15	deer red species found **mouse small** musk white black mule genus
16	peer gynt suite suites hall **op** orchestra mountain king music morning symphony grieg
17	dachshund wiener dog pet racing dogs named american america **lives hot film** dachshunds
18	incandescent light bulb fluorescent filament **led** tungsten lights high **used** electric lighting sources
19	steinbach nutcracker decorative become ulbricht **especially** popular united christian items kolbe **became**
20	mistletoe dwarf known native plant christmas common **new** species genus family
21	mexican food culture cuisine **economic introduced** influenced state foods **many** university california world
22	antique appliance restoration **stream series** appliances cultural **firs**t antiques river work art **national**
23	toronto film awards critics festival annual association **choice given** international city tiff best
24	email benefits business platform marketing messaging high enterprise directory customers company media online
25	composition zirconium **alloys metal compositions nuclear act form hydride lower applications titanium**
26	jennifer aniston stars comedy actress friends rachel jason starring directed role actor
27	royal caribbean cruise lines cruises carnival holland major similar seas islands international **america** largest
28	baltimore city united located orioles major neighborhood maryland county school baseball
29	**delta omega theta phi** chapter pi university **epsilon gamma** fraternity
30	us transportation highway **original** radio results cars **indicate system** reporting
31	babe ruth 1920s yankees baseball home season park yankee **became late** gehrig teams
32	growth rates pine tree high rapid caused **old** pines species vegetation short trunk roots
33	rosebowl parade american channel year series day roses annual film new pasadena
34	auto skoda works wall including xanthi car large **refer simply** manufacturer vehicles
35	gps clock atomic radio signals system navigation **used** receiver positioning satellite **master**
36	eldorado casino reno circus business nevada resort legacy vegas **defunct** rancho casinos hotel
37	angioplasty stent surgery bypass percutaneous **used** stenting treatment artery coronary procedures
38	newport beach california harbor city club balboa united american county **school high located**
39	calcium salt compound acid high channel mineral blood food used carbonate

40	motorcycle safety helmets protective accidents clothing wear seat personal number riders motorcycles equipment
41	japanese wave tsunami **first movement hand published second known time company** waves
42	us savings bonds treasury income credit series social rate bureau tax securities **debt**
43	retire term years forced united **end instead** retirement **first** age year
44	nirvana live rock later buddhism second first death **band written american**
45	decade 1920s music popular era century **part although** decade first period time
46	temporomandibular joint condylar pain disorder mastication muscles articular dysfunction jaw chewing
47	orchid native known plants orchidaceae genus plant name species commonly flowering
48	hair transplant follicular skin transplantation technique loss since called head new transplants
49	pool cue object name term **usually** player billiard small series **hockey another**
50	dna testing genetic forensic analysis tests identification genealogy chromosome fingerprinting**ase** profiling

TABLE 1. ENHANCEMENT TERMS GENERATED WITH IRRELEVANT TERMS HIGHLIGHTED IN BOLD.

Query 41 ('Japanese Wave') produced the terms: tsunami first movement hand published second known time company waves. The ideal mapping would be to /wiki/ The_Great_Wave_off_Kanagawa that represents a specific event in time, however, the vague nature of the query did not provide sufficient context to achieve this mapping.

3.2 Represented Concepts

During the process of gathering and selecting enhancement terms for a given query, the WNSSA gathers a collection of related Wikipedia articles as stem pages or concepts to create a wide collection of a priori knowledge. When this process is in action it can often be the case that the articles that are selected may not be completely relevant to the user's original search query. As a result of this, conceptually distant terms are included during the iterations of the algorithm. Overall, the WNSSA is proficient at identifying these irrelevant terms.

In TABLE 2, a breakdown of the TREC-9 Web topics is shown with the Wikipedia concept mappings identified during the first iteration of the WNSSA. In this table, redirects that were automatically performed by Wikipedia are shown as arrows. This first iteration of the algorithm is the most important as it lays the path for the stem terms that will be used in future iterations. For Query 1 ('What is a Bengals cat?'), the mapping to /wiki/Bengal_cat was correct. In Query 2 ('do beavers live in salt water?') a direct mapping could not be found leaving the algorithm to break the two core concepts of this query part. /wiki/Beaver and /Saline_water are relevant to the query, however, the answer format requested was not fulfilled during retrieval. An issue with the quantity of data returned can be seen with Query 10 ('Who was Moses?') can be seen as although the mapping was correctly made to /wiki/Moses, the history attached to this name caused numerous different representations to be included during enhancement causing the query to fail. The uniqueness of Query 31 ('what did babe ruth do in the 1920's?') provided very little room for error as the 1920s were synonymous with baseball and Babe Ruth. In Query 42 ('us savings bonds') was entered without any additional context, however in Wikipedia the use of article redirects can be seen. In this case /wiki/us_savings_bonds redirected to /wiki/United_States_Treasury_security.

ID	TREC-9 Web Topics	Wikipedia Concepts
1	What is a Bengals cat?	Bengal_cat
2	do beavers live in salt water	Beaver, Salt_water-> Saline_water
3	hunger	Hunger
4	parkinson's disease	Parkinson%27s_disease
5	whan did Jackie Robinson appear at his first game	Jackie_Robinson
6	is the world going to end 2000	Is the world going to end->Eschatology, 2000
7	CHEVROLET TRUCKS	List_of_Chevrolet_vehicles
8	fasting	fasting
9	when can a lender foreclose on property	Predatory_lending, Foreclosed -> Foreclosure
10	Who was Moses?	Moses
11	lava lamps	Lava_lamp
12	real estate and new jersey	Real_estate, New_Jersey
13	tartin	Tartan
14	nativityscenes	Nativity_scene
15	deer	Deer
16	information about the Peer Gynt Suite?	Peer_Gynt, Suite
17	dachshund dachshunds 'wiener dog'	Dachshund, Miniature_Dachshund
18	incandescent light bulb	Incandescent_light_bulb
19	steinbach nutcracker	Nutcracker
20	mistletoe	Mistletoe
21	mexican food culture	Mexican cuisine. Culture of Mexico
22	antique appliance restoration	Antiques restoration, Restoration
23	Toronto FIlm Awards	Toronto Film Critics Association, Toronto International Film Festival
24	how e-mail bennefits businesses	
25	what is the compostion of zirconium	Isotopes of zirconium, Zirconium alloy
26	Jennifer Aniston	Jennifer_Aniston
27	Royal Carribean Cruise Lines	Royal Caribbean International, Royal Caribbean Cruises Ltd.
28	baltimore	Baltimore
29	where can I find information about kappa alpha psi?	Kappa Alpha Psi, Alpha Kappa Psi
30	car traffic report	Trafffic_reporting
31	what did babe ruth do in the 1920's?	The_Babe_Ruth_Story, Babe_Ruth
32	where can i find growth rates for the pine tree?	Dendrochronology
33	rosebowl parade	Rose Parade, Rose Bowl
34	auto skoda	Škoda Auto
35	gps clock	Radio clock, Global Positioning System
36	where is the Eldorado Casino in Reno?	Eldorado Resort Casino
37	angioplast7	Angioplasty
38	newport beach california	Newport Beach, California
39	calcium	Calcium
40	motorcycle safety helmets	Motorcycle helmet
41	Japanese Wave	Japanese New Wave, The Great Wave off Kanagawa
42	us savings bonds	us savings bonds -> United States Treasury security
43	retire	Retire -> Retirement
44	nirvana	Nirvana
45	Where can I find information on the decade of the 1920's?	1920s
46	TMJ	Temporomandibular_joint
47	orchids	Orchids -> Orchidaceae
48	hair transplant	Hair_transplant -> Hair_transplantation
49	pool cue	Pool_cue -> Cue_stick

50	DNA Testing	DNA_testing -> Genetic_testing

TABLE 2. TREC-9 WEB TOPICS WITH WIKIPEDIA ARTICLE CONCEPTS.

4. Discussion

In the analysis that was performed, a direct link to the source data (Wikipedia articles) was used as the stem at the beginning of the WNSSA's first iteration. These early mappings have shown to have a great impact on the final enhanced query result. In some queries, the amount of content stored inside of a Wikipedia article provided difficulty for the algorithm as many concepts were related under one original concept, e.g., Moses, which contained many different historical and people of interest. Often terms entered by a user may be acronyms that may not yet have been published in locations such as static document repositories or thesauri. Wikipedia typically has an advantage at it contains these as they become commonplace in a domain. An example of this can be seen with the acronym 'TMJ' which was then redirect to 'Temporomandibular joint'.

Spelling issues are often remedied during Wikipedia article title redirects. The algorithm did benefit from a spelling correction process on the original test queries. The lack of prior search knowledge during ASQE is still an issue as search queries such as 'Nirvana' can easily be focused on the musicians or the Buddhist state of enlightenment. Small amounts of content about the user or their areas of interest can greatly impact the success of a ASQE algorithm.

The length of the query that was entered can have a great impact on the performance of the algorithm. In a traditional search, the more terms that are added to the query the more focused the search becomes. This can be seen particularly with query 31 ('what did babe ruth do in the 1920's?'). Although the length of the query is longer than a typical query length, the focused nature of the query allowed for a direct mapping to two highly relevant results Wikipedia pages for a priori knowledge, leaving only two generated enhancement terms that could be deemed irrelevant. If longer queries are not focused on one core topic, the enhancement process often produced poor results.

During the analysis of Wikipedia article concepts, less than 1% of the time a direct mapping could not be found. This can be attributed to the fast scope of Wikipedia articles and the utilisations of URL redirects that are common when spelling or alternative titles are used for articles. This provided the algorithm a source of high-quality knowledge before additional iterations of the algorithm were performed. When processing data for use with ASQE algorithms, stop word filtering should be done with care when working with search queries as often the terms that are removed provide the much-needed context for the query.

5. Conclusions

Wikipedia has proven to be a useful resource for ASQE when all the available data and services e.g., backlink API and search functionality is utilised. Although the mapping of queries to Wikipedia articles often provides high-quality enhancement terms, if the search query is short

or contains several stop words the query can easily be reduced to individual tokens without context. This provides an issue for the enhancement algorithm, leaving it to take multiple different directions and possibly following the wrong one. To further aid the enhancement process, additional context of terms relevant to the query could be implemented. Human error e.g., spelling should be corrected before queries are processed by an ASQE algorithm to avoid additional strain on an already difficult process. Any small quantity of context, in terms of a user profile, previous search history or geographic location could further enhance ASQE algorithm results.

The type of query which is entered has a great impact on the enhancement process. If a question format is used, care should be taken to ensure that the phrasing of the question is not used as part of the enhancement process. Wikipedia article redirects has shown to be a beneficial resource during enhancement to map from a user's search query to a relevant article that may have a completely different article title.

References

Asfari, Ounas, Doan, Bich-liên, Bourda, Yolaine and Sansonnet, Jean-Paul. 2009. 'Personalized Access to Information by Query Reformulation Based on the State of the Current Task and User Profile.' Paper presented at Third International Conference on Advances in Semantic Processing, 113-116. IEEE.

Bazzanella, Barbara, Stoermer, Heiko, and Bouquet, Paolo. 2010. 'Searching for individual entities: A query analysis.', Paper presented at International Conference on Information Reuse & Integration, 115-120. IEEE.

Gao, Jianfeng, Xu , Gu and Xu, Jinxi. 2013. Query expansion using path-constrained random walks. Paper presented at 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13), 563-572. ACM.

Goslin, Kyle, Hofmann, Markus. 2017. 'A Comparison of Automatic Search Query Enhancement Algorithms That Utilise Wikipedia as a Source of A Priori Knowledge.' Paper presented at 9th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE'17), 6-13. ACM.

Goslin, Kyle, Hofmann, Markus. 2018. 'A Wikipedia powered state-based approach to automatic search query enhancement.' Journal of Information Processing & Management 54(4), 726-739. Elsevier.

Jansen, Bernard, Spink, Amanda, Bateman, Judy and Saracevic, Tefko. 1998. 'Real life information retrieval: a study of user queries on the Web.' Paper presented at ACM SIGIR Forum 32, 5-17. ACM.

Mastora, Anna, Monopoli, Maria and Kapidakis, Sarantos. 2008. 'Term selection patterns for formulating queries: a User study focused on term semantics.' Paper presented at Third International Conference on Digital Information Management, 125-130. IEEE.

Ogilvie, Paul, Voorhees, Ellen and Callan, Jamie. 2009. 'On the number of terms used in automatic query expansion.' Journal of Information Retrieval 12(6): 666. Springer.

Voorhees, Ellen M. 1994. 'Query expansion using lexical-semantic relations.' Paper presented at the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94), 61-69. Springer-Verlag.