A Tutorial on Information Retrieval Using Query Expansion

Mohamed Yehia Dahab, Sara Alnofaie and Mahmoud Kamel

Abstract Most successful information retrieval techniques which has the ability to expand the original query with additional terms that best represent the actual user need. This tutorial gives an overview of information retrieval models which are based on query expansion along with practical details and description on methods of implementation. Toy examples with data are provided to assist the reader to grasp the main idea behind the query expansion (QE) techniques such as Kullback-Leibler Divergence (KLD) and the candidate expansion terms based on WordNet. The tutorial uses spectral analysis which one of the recent information retrieval techniques that considers the term proximity.

Keywords Information retrieval · Query expansion · Spectral analysis · KLD WordNet

1 Introduction

The most critical issue for information retrieval is the term mismatch problem. One well known method to overcome this limitation is the QE of the original query terms with additional terms that best retrieve the most related documents.

Most successful technique which has the ability to expand the original query with additional words that best capture the actual user goal. Most of the recent information retrieval models are based on the proximity between terms which is useful for

M.Y. Dahab (✉) · S. Alnofaie · M. Kamel
Faculty of Computing and Information Technology, Department of Computing Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
e-mail: mdahab@kau.edu.sa; mohamed.dahab@gmail.com
URL: http://www.kau.edu.sa/Home.aspx

S. Alnofaie
e-mail: salnefaie@kau.edu.sa

M. Kamel
e-mail: miali@kau.edu.sa

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_35

improving document retrieval performance (e.g., [1–8] and [9]), with consideration of the term positional information and different transformation algorithms to calculate the document score. The main idea of the document spectral analysis is extended to be used in document classification and clustering [6] and [9].

In the information retrieval community, Usually, the queries consist of two or three terms, which are sometimes not enough to understand the expectations of the end user and fail to express topic of search. The QE is a process of broadening the query terms using words that share statistical or semantic relationships with query terms in the collection or share meaning with query terms. To our knowledge, there is no technique uses the QE in information retrieval model and considers the proximity information except [7].

This tutorial provides more than a toy example with data to assist the reader to grasp the main idea behind the QE techniques such as KLD and the candidate expansion terms based on WordNet. The tutorial uses spectral analysis which one of the recent information retrieval techniques that considers the term proximity.

This research will proceed as follows, Sect. 2 presents the related background, Sect. 3 lists the preprocessing tasks, Sect. 4 introduces the a brief description about the synthetic dataset used in this research, Sect. 5 demonstrates the application of different techniques of the QE on the dataset and finally the conclusion is given in Sect. 6.

2 Background

Here we review some general aspects that are important for a full understanding of the following tutorials.

2.1 Automatic Query Expansion

The average length of the query is around 2–3 words where may the users and the document collection does not use the same words for the same concept that is known as the vocabulary or mismatch problem. Therefore, there is difficulty in retrieving the relevant documents set. To improve the performance of IR model, use the overcome mismatch problem approaches. One of the successful approaches is to automatically expand the original query with other terms that best capture the actual user intent that makes the query more useful [10] and [11].

Automatic QE process can divided into four steps: data source preprocessing, candidate expansion features generation and ranking, expansion features selection, the query reformulated [12, 13] and [14].

2.2 Data Source Preprocessing

In this step, the data source that is used for expanding the user query transforms into an effective format for the following steps. It consists of two phases. First, extract the intermediate features. Then, construct the appropriate data structures for access and manipulation this features in an easy way. Based on the source of the Candidate Expansion Terms (CET) the QE approach classifies to the external resources such as the WordNet and the target corpus. The WordNet approaches set some or all the synonyms terms of the synset the contain query term as candidate terms. The target corpus approaches are also divided into local and global. The global approaches set the whole corpus terms as candidate terms and analyze it while the local approaches set only the top relevant documents terms of the initial search results. The local approaches are known as pseudo relevance feedback. In the IR model, the documents collection or corpus is indexing to run the query. As seen in the above section, the documents store using inverted index file, which is useful in some QE approach such as the global approach while the local approach needs to the documents using direct index file.

2.3 Features Generation and Ranking

In this stage, the candidate expansion features generate and ranks by the model. The original query and the data source is the input to this stage while the candidate expansion features associated with the scores is the output. A small number of the candidate features add to the query. Therefore, the feature ranking is important. The relationship between the query terms and candidate features classify the generation and ranking approaches to:

A. One-to-one associations.
B. One-to-many associations.
C. Analysis of feature distribution in top-ranked documents.

2.4 Expansion Features Selection

After the candidate features ranking for some QE approach, the limited number of features is added to the query to process the new query rapidly.

2.5 Query Reformulation

This step usually involves assigning a weight to each expansion feature and re-weights each query term before submitting the new query to the IR model. The most popular query re-weighting scheme was proposed in [15].

3 Preprocessing Tasks

Preprocessing Tasks includes all steps to transform the query and the raw data source used for expanding the user query into a format that will be more effectively processed by subsequent steps. This includes the following steps:

- Text extraction from documents like HTML, PDF, MS Word, etc. (if the collection is made of such documents).
- Tokenization (i.e., extraction of individual words, ignoring punctuation and case).
- Stop word removal (i.e., removal of common words such as articles and prepositions).
- Word stemming (i.e., reduction of inflected or derivational words to their root form).
- Word weighting (i.e., assignment of a score that reflects the importance of the word, usually in each document).
- Some Automatic query expansion (AQE) techniques based on system that indexing the document use inverted index file which represent the document as a set of weighted terms. The indexing system may also store term positions, to provide proximity based search. Other AQE techniques, based on corpus analysis, require the extraction of particular features from the collection.
- Creating term signals for each document is an important preprocessing task but it will be shown in the processing phase for the purpose of giving the readers more details and explanation.

4 Synthetic Dataset

The tutorial is based on a single query q which is "*Suicide explosions in city*". Table 1 shows the synthetic dataset used in this research. It contains 20 rows, each row has three columns, the first column exhibits whether the document content is relevant or not (1 means relevant while 0 means not relevant). The query q is applied on the dataset shown in Table 1.

Table 1 Synthetic dataset

Relevant	Document content	Document No.
1	Explosions in Sanaa, city metropolis of Yamen, today. The terrorist group, Houthi organization, claimed the responsibility for the terrorist action. Three blasts in different places. The Downtown bomb left a number of victims	1
1	Burst Two explosions shake Sacramento city which is Located in California. The first one led to the death of five people. The second one resulted in wounding 65 people. Moreover, massive destruction in cars and buildings	2
1	A series of explosions in france. The terrorist Daesh organization, claimed responsibility for Terrorist action in paris city. one bomb burst in the Playground and another bomb in coffee. The detention of a number of hostages in the theater Pataklan	3
1	Burst Two explosions near the Yusufiyah mosque. The terrorist group, Daesh, claimed the responsibility for the accident. This action resulted in wounding two people and killing three. Upon the arrival of paramedics a malefactor bomb himself in the west part of Baghdad city	4
0	A new mall will be opened in Jeddah downtown on 23th of March	5
0	Large number of celebrity gathered in downtown Los Angeles	6
0	Syrians refugees flocking to Europe	7
0	Expo starting in Milano in April 2017	8
0	Snap chat has changed the Privacy Policy	9
0	Houthi organization Announces responsibility for security of Sanaa starting from Feb 2nd 2013	10
0	Arabic speaker passengers were denied boarding American plane	11
0	Terrorist word comes from the Latin word terrorism meaning great fear. Great fear is exactly what they want to create in order to achieve their goals. They are using violence, chaos, bomb, and destruction. By doing this, they are aiming to force people, and governments to take particular action especially for political Economic social and religious purposes. It is highly destructive phenomenon in recent years	12
0	Hundreds of people have demonstrated in Dublin to support anti-racism	13
0	A vigorous car blast in the east part of Beirut that has majority of Christian on Friday	14
1	Terrorist attack by nuclear bomb in Hiroshima Japan by United States of America in August 1945	15
0	Close Brussels Metro after New York state declared a state of emergency to the highest level	16

(continued)

Table 1 (continued)

Relevant	Document content	Document No.
0	More than 160 000 people were the victims of Tsunamis in 26 December 2004. Tsunami is a seismic sea wave. It is a series of waves in a water caused by the displacement of a large volume of water generally in an ocean or a large lake. The region was affected by Tsunami is the metropolis on Indian Ocean	17
1	Many victims in 3 big blasts today evening in Afghanistan metropolis	18
1	In 24th of Mar 2014 morning, Al-Qaeda organization claimed the responsibility for several bursts in Sanaa and Taiz	19
0	President Obama blasts Donald Trump's recent remarks	20

5 The Application of Different QE Techniques

In this section, statistical and semantic QE techniques will be applied using only one of the recent information retrieval techniques which is spectral based information retrieval ([5] and [8]).

5.1 Spectral Based Information Retrieval with QE Using KLD

As shown in Table 2, the matched query terms in document content are underlined. Document score increases as the query terms in documents content close to each other. Document score is computed by applying Haar discrete wavelet transform as explained in details in [2, 5] and [8]. Rows from 5 to 20 are empty because they do not have a matched query term. The notion of term signal introduced by [1–3, 8] is a vector representation of terms that describes frequencies of term occurrences in particular partitions within a document. In the example, 8 partitions or bins are used. Each line in the document content represents a bin, that is why there are 8 lines. The term signal of the term "*Suicide*" has been neglected because there is no term matched with it. The underline term means it matches with one of query terms. Some important preprocessing have been applied on document content. Documents number 15, 18 and 19 are relevant but they have zero document score so they have not been retrieved. By computing the Mean Average Precision (MAP) defined by the following equation, $MAP(q) = \frac{1}{7}(1 + 1 + 1 + 1 + 0 + 0 + 0) \approx 0.57$.

$$MAP(q) = \frac{1}{N}\Sigma_{j=1}^{N}\frac{1}{Q}\Sigma_{i=1}^{Q}P(doc_i) \tag{1}$$

Table 2 Document scores using spectral based information retrieval with the original query

No.	Document content	Term signal	Document score
1	– Explosions sanaa city – Metropolis yamen today – Terrorist group houthi – Organization claimed – Responsibility terrorist action – Blasts places downtown – Bomb left – Number victims	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [1, 0, 0, 0, 0, 0, 0, 0]	1.778
2	– Burst explosions shake – Sacramento city – Located california – Led death – People wounding – 65 people – Massive destruction – Cars buildings	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 1, 0, 0, 0, 0, 0, 0]	0.889
3	– Series explosions france – Terrorist daesh organization – Claimed responsibility terrorist – Action paris city – Bomb burst playground – Bomb coffee detention – Number hostages – Theater pataklan	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 1, 0, 0, 0, 0]	0.556
4	– Burst explosions yusufiyah – Mosque terrorist group – Daesh claimed responsibility – Accident action wounding – People killing arrival – Paramedics malefactor – Bomb west – Baghdad city	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 1]	0.389
5			0.0
6			0.0
7			0.0
8			0.0
9			0.0
10			0.0
11			0.0
12			0.0
13			0.0
14			0.0

(continued)

Table 2 (continued)

No.	Document content	Term signal	Document score
15			0.0
16			0.0
17			0.0
18			0.0
19			0.0
20			0.0

Where:

N: is number of queries.
Q: is number of relevant documents for query q.
$P(doc_i)$: is the precision of ith relevant document.

To distinguish between useful candidate expansion term and unuseful expansion term by comparing the distribution of this term in the top relevant documents of the query with the distribution of this term in all documents. In other words, the score of the appropriate expansion term is high when the percentage of this term appearance in relevant documents more than in the collection.

Computing KLD. Carpineto proposed interesting query expansion approaches based on term distribution analysis [10]. The distributions variance between the terms in the top relevant documents and entire document collection where those terms obtain from the first pass retrieval using the query. The query expands with terms that have a high probability in the top related document compare with low probability in the whole set. The KLD score of term in the CET are compute using the following equation:

$$KLDScore(t) = P_R(t) log \frac{P_R(t)}{P_C(t)} \tag{2}$$

where $P_R(t)$ is the probability of the term t in the top ranked documents R, and $P_C(t)$ is the term t probability in the corpus C, given by the following equations:

$$P_R(t) = \frac{\Sigma_{d \in R} f(t, d)}{\Sigma_{d \in R} |d|} \tag{3}$$

$$P_C(t) = \frac{\Sigma_{d \in C} f(t, d)}{\Sigma_{d \in C} |d|} \tag{4}$$

After sorting the documents according to document scores to form the pseudo documents. Let the number of pseudo documents ($k = 4$) which includes the documents number 1, 2, 3 and 4 respectively. Apply the previous three equations on each term pseudo documents.

Let $NT_R = \Sigma_{d \in R}|d|$ and $NT_C = \Sigma_{d \in C}|d|$.
Now $NT_R = 80$ and $NT_C = 225$. $\Sigma_{d \in R} f(malefactor, d) = 1$.
Also $\Sigma_{d \in C} f(malefactor, d) = 1$.

To compute the KLD score of the term *malefactor*, $P_R(malefactor) = \frac{1}{80} = 0.0125$ and $P_C(malefactor) = \frac{1}{225} = 0.0044$.
$KLDScore(malefactor) = P_R(malefactor) log \frac{P_R(malefactor)}{P_C(malefactor)} \approx 0.019$.
The KLD score for each term in pseudo documents show in Table 3.

Table 3 KLD score for each term in pseudo documents

Term	KLD score	Term	KLD score
City	0.075	Today	0.019
Explosions	0.075	Malefactor	0.019
Terrorist	0.063	Coffee	0.019
Bomb	0.045	Sacramento	0.019
Claimed	0.03	Detention	0.019
Burst	0.03	Baghdad	0.019
Action	0.03	California	0.019
Group	0.027	Hostages	0.019
Wounding	0.027	Buildings	0.019
Daesh	0.027	Killing	0.019
Responsibility	0.021	Accident	0.019
Number	0.017	Theater	0.019
People	0.015	Places	0.019
Arrival	0.019	Pataklan	0.019
Yusufiyah	0.019	Massive	0.019
65	0.019	West	0.019
Shake	0.019	Left	0.019
Led	0.019	Organization	0.01
Death	0.019	Series	0.005
Paris	0.019	Destruction	0.005
Paramedics	0.019	Houthi	0.005
Mosque	0.019	Blasts	0.005
France	0.019	Metropolis	0.0
Located	0.019	Victims	0.0
Playground	0.019	Downtown	0.0
Cars	0.019	Sanaa	0.0
Yamen	0.019		

Suppose the maximum length of candidate term list is four, so select the four terms that have maximum KLD score from Table 3. The candidate term list includes city, explosions, bomb and terrorist.

The candidate term list includes includes the original query terms in addition to extra terms that are supposed to be related to the original terms.

Apply the information retrieval technique, which was the spectral based information retrieval, again on the new query contains the new candidate term list.

Table 4 shows the document scores using spectral based information retrieval with expansion query using KLD. Documents, that do not have score, have been removed from table. The weight of original query terms has been increased by 100% to magnify their contribution in the document score more than the expanded query terms.

As a results from increasing the candidate term list to four, documents 12 and 15 are retrieved and it affects on MAP as following:

$MAP(q) = \frac{1}{7}(1 + 1 + 1 + 1 + 0.83 + 0 + 0) \approx 0.69.$

MAP as information retrieval measure has been improved when using statistical query expansion. Also, the order of documents according to document scores becomes 3, 1, 4, 2, 15 and 12 respectively.

5.2 Spectral Based Information Retrieval with QE Using WordNet

To apply spectral based information retrieval using semantic QE approach with the semantic lexicon WordNet, the following steps should be carried out:

1. **Determining the number of both pseudo documents and candidate term list**
   To continue using the same data described in the previous subsection, the number of pseudo documents is four, that is also include documents from $1 - 4$. The top two related terms will be added to the candidate term list.
2. **Computing the semantic similarity**
   For each term, t, in pseudo documents and term q_i in the query, compute related score using the following formulas:

   (a) Compute the semantic similarity between the term t and q_i using WordNet by considering the definitions of t and q_i as two sets of words, and the overlap between these two sets is taken as $Rel(t, q_i)$.

$$Rel(t, q_i) = \frac{2 * C_{t,q_i}}{C_t + C_{q_i}} \tag{5}$$

   where c_t, c_{q_i} is the number of words in t, q_i definitions respectively.
   C_{t,q_i} is the number of common words.
   To compute $Rel(t, q_i)$ when t = metropolis and q_i = city. The definition of the term "*metropolis*" in WordNet is :

Table 4 The document scores using spectral based information retrieval with the expansion query using KLD

No.	Document content	Term signal	Document score
1	– Explosions sanaa city – Metropolis yamen today – Terrorist group houthi – Organization claimed – Responsibility terrorist action – Blasts places downtown – Bomb left – Number victims	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [1, 0, 0, 0, 0, 0, 0, 0] Terrorist [0, 0, 1, 0, 1, 0, 0, 0] Bomb [0, 0, 0, 0, 0, 0, 1, 0]	5.635
2	– Burst explosions shake – Sacramento city – Located california – Led death – People wounding – 65 people – Massive destruction – Cars buildings	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 1, 0, 0, 0, 0, 0, 0] Terrorist [0, 0, 0, 0, 0, 0, 0, 0] Bomb [0, 0, 0, 0, 0, 0, 0, 0]	1.28
3	– Series explosions france – Terrorist daesh organization – Claimed responsibility terrorist – Action paris city – Bomb burst playground – Bomb coffee detention – Number hostages – Theater pataklan	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 1, 0, 0, 0, 0] Terrorist [0, 1, 1, 0, 0, 0, 0, 0] Bomb [0, 0, 0, 0, 1, 1, 0, 0]	6.44
4	– Burst explosions yusufiyah – Mosque terrorist group – Daesh claimed responsibility – Accident action wounding – People killing arrival – Paramedics malefactor – Bomb west – Baghdad city	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 1] Terrorist [0, 1, 0, 0, 0, 0, 0, 0] Bomb [0, 0, 0, 0, 0, 0, 1, 0]	3.6
12	– Terrorist word latin word – Terrorism meaning great fear – Great fear create order – Achieve goals bomb violence – Chaos destruction aiming force – People governments action political – Economic social religious purposes – Highly destructive phenomenon years	Explosions [0, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 0] Terrorist [1, 0, 0, 0, 0, 0, 0, 0] Bomb [0, 0, 0, 1, 0, 0, 0, 0]	0.2
15	– Terrorist attack – Nuclear bomb – Hiroshima – Japan – United – America – August – 1945	Explosions [0, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 0] Terrorist [1, 0, 0, 0, 0, 0, 0, 0] Bomb [0, 1, 0, 0, 0, 0, 0, 0]	0.32

> a large and densely populated urban area; may include several independent administrative districts people living in a large densely populated municipality metropolis

The definition of the term "*city*" in WordNet is:

> a large and densely populated urban area; may include several independent administrative districts an incorporated administrative district established by state charter people living in a large densely populated municipality city

$C_t = C_{metropolis} = 22$, $C_{q_i} = C_{city} = 31$, $C_{t,q_i} = C_{metropolis,city} = 18$.
$Rel(metropolis, city) = \frac{2*18}{31+22} = 0.679$.

(b) Compute idf_t for each term in the collection.
$idf_{metropolis} = 0.673$.

(c) Select the most relevant document of the pseudo documents in which t occurs. This is intended to capture the intuition that terms coming from relevant document are better than the terms coming from non-relevant documents.

$$S(t, q_i) = Rel(t, q_i) * idf_t * \Sigma_{(d \in R)} \frac{sim(d, q)}{max_{(d' \in R)} sim(d', q)} \tag{6}$$

The term "metropolis" exist in only one pseudo documents 1 with score 1.778.
$Sim(metropolis) = \frac{1.778}{1.778} = 1$.

(d) The relatedness score of t with the whole documents is given by

$$S(t) = \Sigma_{q_i \in q} \frac{S(t, q_i)}{1 + S(t, q_i)} \tag{7}$$

$S(metropolis, city) = Rel(metropolis, city) * idf_{metropolis} * Sim(metropolis)$
$= 0.679 * 0.673 * 1 = 0.457$.
compute $S(metropolis, explosions)$ and $S(metropolis, Suicide)$ to compute $S(metropolis)$.

3. **Selecting the best related score**
The relatedness score of the remaining terms in pseudo documents shown in Table 5.

Table 5 The relatedness score of the terms in pseudo documents

Term	Relatedness score	Term	Relatedness score
Explosions	0.52	Detention	0.084
City	0.52	Malefactor	0.082
Metropolis	0.377	Playground	0.08
Victims	0.252	Mosque	0.067
Organization	0.23	Paramedics	0.065
Sanaa	0.229	Killing	0.065
Group	0.214	Series	0.061
Downtown	0.207	Accident	0.05
Action	0.203	West	0.041
Blasts	0.187	People	0.035
Left	0.182	Sacramento	0.027
Today	0.176	Located	0.027
Responsibility	0.165	Death	0.026
Claimed	0.164	Cars	0.024
Number	0.159	California	0.023
Terrorist	0.152	Massive	0.023
Bomb	0.148	Destruction	0.02
Places	0.139	Buildings	0.016
France	0.129	Shake	0.014
Arrival	0.123	Led	0.01
Paris	0.115	Yusufiyah	0.0
Theater	0.113	65	0.0
Burst	0.111	Daesh	0.0
Coffee	0.107	Yamen	0.0
Hostages	0.099	Houthi	0.0
Wounding	0.091	Pataklan	0.0
Baghdad	0.085		

4. **Applying spectral based information retrieval with new expansion query using WordNet**
   The top related terms that will be added to the candidate term list are Metropolis and victims. Table 6 shows the results of applying the spectral based information retrieval with new expansion query using WordNet.
   As shown in the Table 6, in addition to the original retrieved documents, i.e. the pseudo documents, documents 17 and 18 have been retrieved. Note that document number 17 is irrelevant while document number 18 is relevant.
   Addition weight, 100%, has been added to the original query terms because they contribute more in the document score. The underline term means it matches with one of query terms.

Table 6 The document scores using spectral based information retrieval with the expansion query using WordNet

No.	Document content	Term signal	Document score
1	– Explosions sanaa city – Metropolis yamen today – Terrorist group houthi – Organization claimed – Responsibility terrorist action – Blasts places downtown – Bomb left – Number victims	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [1, 0, 0, 0, 0, 0, 0, 0] Metropolis [0, 1, 0, 0, 0, 0, 0, 0] Victims [0, 0, 0, 0, 0, 0, 0, 1]	6.38
2	– Burst explosions shake – Sacramento city – Located california – Led death – People wounding – 65 people – Massive destruction – Cars buildings	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 1, 0, 0, 0, 0, 0, 0] Metropolis [0, 0, 0, 0, 0, 0, 0, 0] Victims [0, 0, 0, 0, 0, 0, 0, 0]	1.28
3	– Series explosions france – Terrorist daesh organization – Claimed responsibility terrorist – Action paris city – Bomb burst playground – Bomb coffee detention – Number hostages – Theater pataklan	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 1, 0, 0, 0, 0] Metropolis [0, 0, 0, 0, 0, 0, 0, 0] Victims [0, 0, 0, 0, 0, 0, 0, 0]	0.8
4	– Burst explosions yusufiyah – Mosque terrorist group – Daesh claimed responsibility – Accident action wounding – People killing arrival – Paramedics malefactor – Bomb west – Baghdad city	Explosions [1, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 1] Metropolis [0, 0, 0, 0, 0, 0, 0, 0] Victims [0, 0, 0, 0, 0, 0, 0, 0]	0.56
17	– 160000 people victims – Tsunamis 26 december 2004 – Tsunami seismic sea wave – Series waves water caused – Displacement large volume water – Generally ocean large – Lake region tsunami – Metropolis indian ocean	Explosions [0, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 0] Metropolis [0, 0, 0, 0, 0, 0, 0, 1] Victims [1, 0, 0, 0, 0, 0, 0, 0]	0.14
18	– Victims – 3 – Big – Blasts – Today – Evening – Afghanistan – Metropolis	Explosions [0, 0, 0, 0, 0, 0, 0, 0] City [0, 0, 0, 0, 0, 0, 0, 0] Metropolis [0, 0, 0, 0, 0, 0, 0, 1] Victims [1, 0, 0, 0, 0, 0, 0, 0]	0.14

Also in Table 6, 8 partitions or bins are used. Each line in the document content represents a bin, that is why there are 8 lines. The term signal of the term "*Suicide*" has been neglected because there is no term matched with it.

5. **Sorting documents based on score**
   After sorting the documents according to the document score, documents 1, 2, 3, 4, 17 and 18 have the following score 6.38, 1.28, 0.8, 0.56, 0.14 and 0.14 respectively.

6. **Evaluation**
   As a results from increasing the candidate term list to four, documents 17 and 18 are retrieved and it affects on MAP as following:
   $MAP(q) = \frac{1}{7}(1 + 1 + 1 + 1 + 0.83 + 0 + 0) \approx 0.69$.
   MAP as information retrieval measure has been improved when using WordNet query expansion.

6 Conclusion

This tutorial shows the impact of extending the query by adding statistical and semantic related terms to the original query terms over proximity based IR system. This is done by combining the spectral based information retrieval model with the best QE approaches such as the distribution approach (KLD) and WordNet. The toy examples results show that the spectral based information retrieval with QE using KLD and WordNet outperformed the spectral based information retrieval in precision at top documents and MAP metric. The toy examples provided in this research demonstrates that dividing the documents into a specific number of segments (8 bin).

References

1. Palaniswami, M., Ramamohanarao, K., Park, L.: Fourier domain scoring: a novel document ranking method. IEEE Trans. Knowl. Data Eng. **16**(5), 529539 (2004)
2. Park, L.A.F., Ramamohanarao, K., Palaniswami, M.: A novel document retrieval method using the discrete wavelet transform. ACM Trans. Inf. Syst. (TOIS). pp. 267–298 (2005)
3. Park, L.A.F., Palaniswami, M., Ramamohanarao, K.: Internet documentltering using fourier domain scoring. In: de Raedt, L., Siebes, A. (Eds.) Principles of Data Mining and Knowledge Discovery, September 2001, number 2168 in Lecture Notes in Articial Intelligence, pp. 362–373. Springer-Verlag (2001)
4. Park, L.A.F., Palaniswami, M., Ramamohanarao, K.: A novel document ranking method using the discrete cosine transform. IEEE Trans. Patt. Analys. Mach. Intell. pp. 130–135 (2005)
5. Aljaloud, H., Dahab, M., Kamal, M.: Stemmer impact on Guranic mobile information retrieval performance. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **7**(12), 135–139 (2016). https://doi.org/10.14569/IJACSA.2016.071218

6. Al-Mofareji, H., Kamel, M., Dahab, M.Y.: WeDoCWT: a new method for web document clustering using discrete wavelet transforms. J. Inf. Knowl. Manage. **16**(1), 1–19 (2017). https://doi.org/10.1142/S0219649217500046
7. Alnofaie, S., Dahab, M., Kamal, M.: A novel information retrieval approach using query expansion and spectral-based. Int. J. Adv. Comput. Sci. Appl. **7**(9), 364–373 (2016). https://doi.org/10.14569/IJACSA.2016.070950
8. Dahab, M.Y., Alnofaie, S., Kamel, M.: Further investigations for documents information retrieval based on DWT. In: Hassanien, S.K.A. (Ed.), International Conference on Advanced Intelligent Systems and Informatics, vol. 533, pp. 3–11. Springer, Cairo (2016). https://doi.org/10.1007/978-3-319-48308-5_1
9. Diwali, A., Kamel, M., Dahab, M.: Arabic text-based chat topic classification using discrete wavelet transform. Int. J. Comput. Sci. **12**(2), 86–94 (2015). Retrieved from http://www.ijcsi.org/papers/IJCSI-12-2-86-94.pdf
10. Kakde, Y.: A Survey of Query Expansion Until. Indian Institute of Technology, Bombay (2012)
11. Singh, J., Sharan, A., Siddiqi, S.: A literature survey on automatic query expansion for effective retrieval task. Int. J. Adv. Comput. Res. **3**(3), 170–178 (2013)
12. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. (CSUR) **44**(1), 1–50 (2012)
13. Ooi, J., Ma, X., Qin, H., Liew, S.C.: A survey of query expansion, query suggestion and query refinement techniques. In: Proceedings of the International Conference on Software Engineering and Computer Systems, pp. 112–117. IEEE (2015)
14. Rocchio, J.J.: Relevance feedback in information retrieval. In: Proceedings of the SMART Retrieval System-Experiments in Automatic Document, pp. 313–323 (1971)
15. Carpineto, C., De Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. ACM Trans. Inf. Syst. (TOIS) **19**(1), 1–27 (2001)