International Journal of Geo-Information

MDPI

Article

Collecting Typhoon Disaster Information from Twitter Based on Query Expansion

Zi Chen * and Samsung Lim

School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia; s.lim@unsw.edu.au

* Correspondence: zi.chen1@student.unsw.edu.au

Received: 2 February 2018; Accepted: 29 March 2018; Published: 2 April 2018

Abstract: Social media is a popular source of volunteered geographic information owing to its massive real-time data; however, the use of social media data in the context of geospatial analysis is challenging because complex semantic filters are required for the aggregation of geographic messages from the data streams. This article proposes a new query expansion method for social media streams which updates the query keywords periodically by the words extracted from the preceding search results. The proposed method has optimized the trade-off between precision and coverage of geographical messages by factoring in the influences of the keyword number and refresh cycle in the query process, and some improvements on the classic Term Frequency-Inverse Document Frequency (TF-IDF) method for short texts were achieved. Furthermore, a number of filters based upon relevance to the target topic were established and tested. This method was tested on a dataset from Twitter within the geographic extent of Macau in August 2017 during two consecutive typhoon hits. The result supports its effectiveness with a controllable precision and considerable increment of relevant information. Moreover, the query keywords can adjust themselves to the local language environment by discovering new keywords. To conclude, this query expansion method is able to provide a reliable method for social media-based information retrieval.

Keywords: information retrieval; social media; typhoon; query expansion

1. Introduction

Volunteered geographic information (VGI) [1] has expanded the sources of geographic information from experts to general public, and shifted geographic information systems (GIS) from an abstruse technology to a medium of communication [2]. Social media is one of the most popular channels where numerous users actively share explicit or implicit geographic information every day. Social media is a novel information source in solving complex problems, such as disaster management, which requires rapid situation awareness and decision-making [3,4], and many efforts have been made on the interpretation of these informal, noisy messages into concise and comprehensible outcomes such as a unified map [5,6]. However, only few studies have focused on data acquisition approaches that could vastly affect the quality of the geographic analysis, probably because they are defined and confined in advance by social media application programming interfaces (APIs).

In most studies, the process of social media data collection consists in searching by a couple of fixed keywords or hashtags to filter relevant information [7–10]. These keywords or hashtags, which normally define or summarize the topic or incident, are usually manually selected to ensure a high precision of search results and remain unchanged throughout the search process. However, this intuitive approach may cause omission of a massive amount of relevant information if the quantity of the keywords is limited or the keywords are not appropriately chosen, which leads to incomplete information retrieval. Furthermore, as the discussion of the topic develops on social media, the

discussion focus may deviate from the initial keyword set, or the keywords become too general to cover a varying corpus of the topic, particularly in a different language environment; as a result, a high volume of irrelevant information can be collected.

Nonetheless, the cure for these problems in data collection of developing streams has long existed in information retrieval, known as query expansion or reformulation. It renovates the original query by updating the query keywords or their weights to enhance efficiency of information retrieval [11]. According to different sources of the new keywords, available techniques in query expansion can be divided into two classes: global and local analysis [12,13]. Global analysis utilizes external corpus or thesaurus, a vocabulary recording terms, and their relationships. In a typical case, the thesaurus will return synonyms and other relevant terms to update the query. Global analysis requires no user input but causes high consumption of computing resources on the establishment and statistics of the massive thesaurus [12,14]. In local analysis, the materials for keyword updates are relevance feedback, which is the relevance evaluation from users on historical search results. It avoids extensive work on thesaurus but encounters negative response from users in practice. To further spare the manual process of user evaluation, the method of blind relevance feedback automatically assumes that the top k results returned by query are relevant, although it is less dependable.

Some studies have introduced query expansion into the acquisition of social media data. It is also known as topic tracking. Different methods are utilized in its implementation. For example [15], modified the Boolean query on Twitter manually as the conversation develops. Certain flexibility can be gained in this approach, but manual operation and the training cost also impose restrictions on search scale and coverage. The authors of [16–19] proposed probabilistic models for topic transition combined with co-occurrence of terms, recency, or/and social relationship, and produced new keywords by parameter optimization. The authors of [20] used information from Web and British Broadcasting Corporation (BBC) news to improve queries. However, there are problems with the aforementioned methods. Firstly, in most cases of social media data collection the coverage of query keywords or recall was not discussed. Secondly, the influences of fundamental parameters such as time interval of query expansion and amount of query keywords were not investigated. The latter especially was confined to a small number without confirming its adequacy. Thirdly, vocabulary mismatch in social media was not addressed. Most global social media platforms are multilingual, which requires more flexibility on query expansion in addition to misspelling and syntax errors. Finally, the classic blind relevance feedback method known as Term Frequency-Inverse Document Frequency (TF-IDF) weighting is rarely leveraged on social media data collection in contrast to its extensive use in information retrieval, probably because the brevity of social media messages may affect its performance [20].

In this paper, a query expansion method based on TF-IDF is proposed to cope with the dynamics of social media streams and enhance coverage in geographic information retrieval from social media while a high-level precision can be guaranteed. All messages were deduplicated and translated into English first in pre-processing. To improve the performance of TF-IDF, short texts were connected to constitute a longer document, and filters based on term co-occurrence were applied to screen both keywords and social media messages. Furthermore, we investigated the influences of parameters and filters to optimize the coverage and precision trade-off. In addition, to verify the feasibility of the proposed search method, experiments were conducted on Twitter data to extract typhoon-related geographic information from the study area of Macau, which was severely struck in August 2017 by two consecutive typhoons, Hato and Pakhar.

2. Materials and Methods

2.1. Data Collection and Pre-Processing

Dataset processed in the experiment was collected through the Twitter Search API from 4:00 p.m. on 21 August 2017 to 3:59 p.m. on 30 August 2017 with a geographical location where the study area contains the administrative region of Macau defined by a minimum circle at coordinates 22.163

and 113.57 and a radius of 10 km. The *geofence* search returns both geolocated tweets and tweets by users with location information in their profile that falls within the search region. Since substantial reduplicative messages severely affect the frequencies of specific terms and TF-IDF weights, duplicate or nearly identical texts were removed from 118,880 messages acquired in the period.

The similarity of two tweets is calculated by Equation (1).

$$\text{Similarity} = 2 * \mathbf{m}/\mathbf{n} \quad (1)$$

where **n** is the total number of words in two tweets, and **m** is the number of identical words of tweets.

Similarity in Equation (1) is implemented by function SequenceMatcher.ratio() from Python package *difflib*. When Similarity exceeds the threshold of 0.9, two tweets are recognized as nearly identical, and the one published at a later time will be deleted. After deduplication, 27,665 messages remain. Afterwards, the original tweets are translated into English from more than six languages including Chinese, Portuguese, Japanese, Korean, Hindi, and Arabic with Google Translate, facilitated by the language label of the tweets. In the process of translation, discernible errors are unavoidable, especially for one message in multiple languages.

Apart from social media messages, 9 news articles published between 21 August 2017 and 23 August 2017 from the website of the local news media Macau Daily Times were exported as a corpus. These news articles consisting of typhoon forecasts were used to imitate the regional discussion in the context of approaching typhoon. Twelve initial keywords, listed in Section 3.3, were extracted from the articles about typhoons by TF-IDF weighting and selected manually afterwards in consideration that these query keywords applied in the first round will significantly affect the search results and query evolution.

2.2. Search Strategy and Initial Parameters

The search process is depicted in Figure 1 below. The dataset of tweets is divided into equivalent time spans and retrieved successively to simulate real-time query. Twelve initial keywords acquired in the pre-processing phase are leveraged to collect relevant messages in the first round. In the midst of social media streams, messages that contain at least one keyword and screened by numerical filters are regarded as the search results. Among the search results, items are labelled positive if they are recognized as related to typhoon, and negative otherwise.

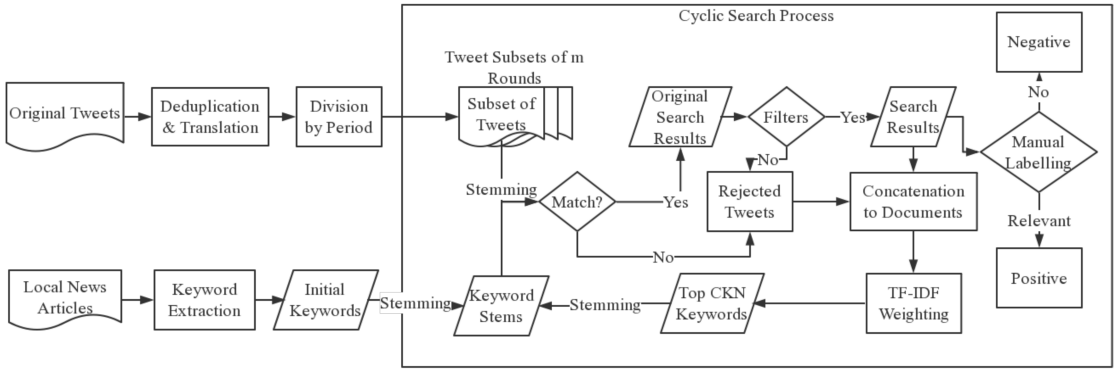


Figure 1. Search process.

To regulate the cyclical process, two initial parameters in the search strategy are defined as follows:

Period refers to the time interval between two keyword updates. Messages collected in one period will be the materials of TF-IDF weighting that form the query keywords in the next period. TF-IDF is a method to measure the representativeness of a word in a document. It consists of two elements: Term Frequency (TF) and Inverse Document Frequency (IDF). The former is the number of occurrences of a

term in a document after normalization. However, as an individual index, it easily falls into the trap of over-counting common words such as 'the'. Therefore, IDF, which quantifies the universality of the term among documents, is introduced. The calculation of TF-IDF for a term is defined in Equation (2).

$$\mathrm{TF-IDF}_{k,j} = \mathrm{TF}_{k,j} * \mathrm{IDF}_{k,j} = \frac{f_{t_k,d_j}}{\sum_i f_{t_i,d_j}} * \log \frac{|D|}{1+|\{d_n \in D : t_k \in d_n\}|} \quad (2)$$

where t_k is the term of which the TF-IDF weight is to be calculated in document d_j; $f_{t,d}$ is the number of occurrences of the term t in the document d; $|D|$ means the amount of documents in corpus D; $|\{d_n \in D : t_k \in d_n\}|$ is the number of documents that contain the term t_k, and 1 is added to this denominator to avoid division by 0. In this paper, TF-IDF weights are computed with TfidfTransformer from the Python package sklearn.feature_extraction.text.

The second initial parameter is the Candidate Keyword Number (CKN). At the end of every period, new keywords are extracted from the search results by ranking of TF-IDF weights and utilized for searching in the next round. These new keywords are named Candidate Keywords because further filtering is essential to refine the keywords.

Moreover, to enhance precision and integrity of the search results, some alterations have been made to the search process.

Search results are concatenated into a single text in each round for TF-IDF weighting because most social media messages are too short for selecting representative keywords. Suppose that a collection of all tweets is T consisting of search results R and rejected messages N. Cardinality of N normally outclasses that of T owing to the sparsity of relevant information in Twitter streams. All messages in R will compose an individual document d_r, and tweets in N analogously form documents $D_n = \{d_1, d_2, \ldots, d_n\}$ with the same tweet number $|R|$ for each document except d_n because of possible remainder. D_n and d_r comprise the corpus for TF-IDF weighting where new keywords are derived from d_r.

In addition, to reduce common words in search keywords, a stop words list that defines the keywords to be deprecated was created. Finally, every keyword is transformed into a word stem to expand the search and matched with word stems from the message to determine whether the message is relevant to the topic or not. The stemmer used in this research is nltk.stem.lancaster from package *nltk* based on the Lancaster stemming algorithm.

2.3. Filters

Even though TF-IDF is a widespread and eminent approach to extract keywords from texts, its outputs may still be irrelevant and lead to deviation from the topic. To control the query evolution, screening on search results is indispensable, and it is two hypotheses that can be the foundation of numerical filters. First is the co-occurrence with the core keyword. If a keyword frequently occurs along with the core word 'typhoon' or a message contains 'typhoon', it is more likely to be highly relevant. The first indicator derived from this hypothesis is defined as OverlapRate in Equation (3).

$$\text{Overlap Rate} = \frac{|S(K) \cap S('\text{typhoon}')|}{|S(K)|} \quad (3)$$

where K is the keyword to be filtered, and S is search results by K. The adoption of this indicator means 'typhoon' will be a constant keyword in the search process.

The second hypothesis is the co-occurrence with other candidate keywords. A message is more likely to be relevant if it contains multiple candidate keywords. Similar conclusions can be inferred for a keyword if it repeatedly appears with other candidate keywords. This hypothesis is embodied as

another indicator, Single Keyword Message Ratio (SKMR), that represents the ratio of messages with merely one keyword as defined by Equation (4).

$$\text{SKMR} = \frac{|\text{SS(K)}|}{|\text{(S(K)}|} \tag{4}$$

where K is the keyword to be filtered, S is the search results by K, and SS is a subset of S in which messages include only one candidate keyword K.

Furthermore, filter strategies are divided into overall and partial removals on the basis of the two indicators. This division is designed on the basis of the second hypothesis for more exquisite filtering. Overall removal abandons the keyword itself and all its search results if the indicator value exceeds the threshold, while the latter deletes simple search results with only one keyword unless the keyword is 'typhoon'. Integrated with these indicators, the following four filters are generated:

Overlap Rate Overall Limit (OROL): The keyword and its search results will be deleted if its Overlap Rate is equal to or less than the threshold.

Overlap Rate Partial Limit (ORPL): Search results with merely one keyword will be deleted if the Overlap Rate of the keyword is equal to or less than the threshold.

Single Keyword Message Ratio Overall Limit (SKMR OL): The keyword and its search results will be deleted if the SKMR of the keyword is equal to or greater than the threshold.

Single Keyword Message Ratio Partial Limit (SKMR PL): Search results with a single keyword will be deleted if the SKMR of the keyword is equal to or greater than the threshold.

The filters, or binary classifiers, can be leveraged separately or conjunctively, depending on their classification precision, which is the ratio of positive messages to search results. Coverage, which describes how all the relevant items in the stream are included in the search results, is measured by the number of positive messages. In the following experiment, the validity of two hypotheses and four filters will be verified.

3. Results

As the framework and parameters of the search method were finalized, the effects of initial parameters and filters were tested in experiments for best values of retrieval effectiveness, namely the optimization of trade-off between precision and coverage.

3.1. Influence of Initial Parameters

3.1.1. Period

Different values of Period were attempted under various CKN. As shown in Figure 2, unstable uptrends of precision emerged along with the increase in Period, and the fluctuation stabilized with the increase in CKN. In general, positive messages substantially declined as Period increased except for search results at CKN = 50. This implies that a shorter Period contributes to more positive results. A Period of 12 h was examined likewise, but no results were retrieved in some rounds. Therefore, 24 h is the optimal option for the parameter Period in the tested values.

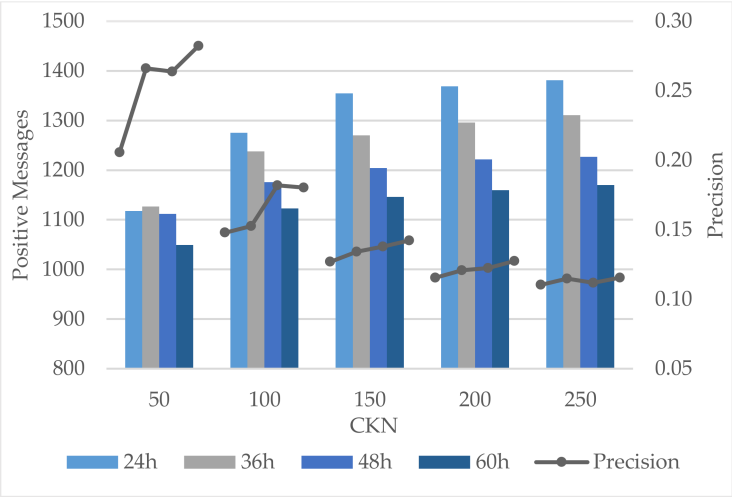


Figure 2. Influence of Period values.

3.1.2. Candidate Keyword Number (CKN)

In this section, the effects of CKN are examined. For the Period set at 24 h, search results, comprised of negative and positive messages, are displayed at a CKN from 50 to 250 in Figure 3. The increase in CKN leads to the growth of both search results and positive messages, but this tendency diminishes as the overlap of search results retrieved by individual keywords increases.

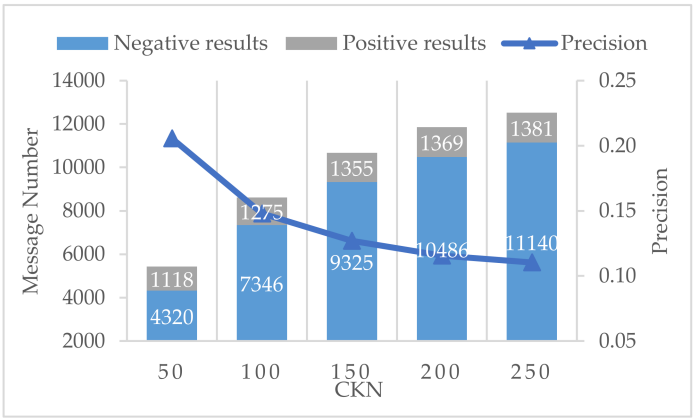


Figure 3. Influence of CKN values.

The increase in CKN gradually causes precision degradation, which is clear in Figure 3; this supports the validity of TF-IDF-weight-based ranking by revealing that keywords of lower ranking produce a lower search precision. For this reason, further filters are essential since a search strategy without filters results in a poor precision around 0.15 no matter what CKN value is chosen. The increase in positive messages is negligible compared to the exploding negatives when CKN > 150, but the highest CKN (250) is chosen because filters can be used later to reduce false positives.

3.2. Filter Effects

3.2.1. Overlap Rate Overall Limit (OROL)

Diverse values of OROL were investigated with the Period fixed at 24 h and CKN = 250, with other filters inactive. The outcomes are plotted in Figure 4. An OROL of −1 refers to no filter involved. The increase in OROL causes a decrease in positive messages. The decrement of positive messages

becomes significant after OROL reaches 0.05. This result implies that OROL is a strong and sensitive constraint on precision even with small changes in its value.

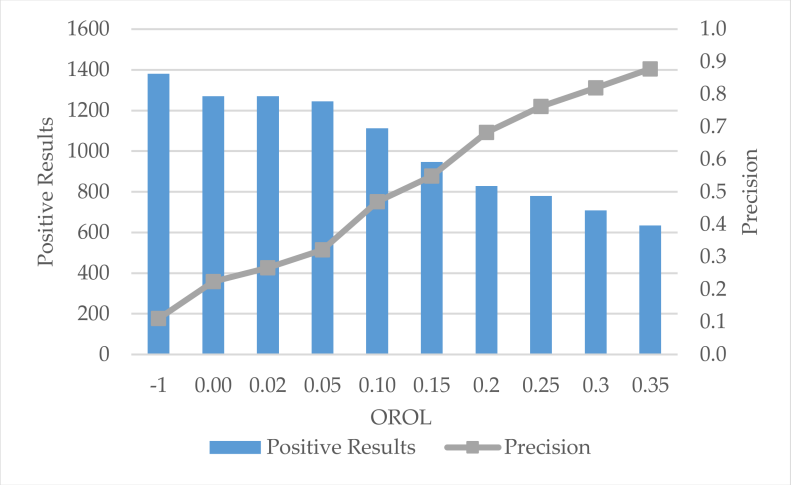
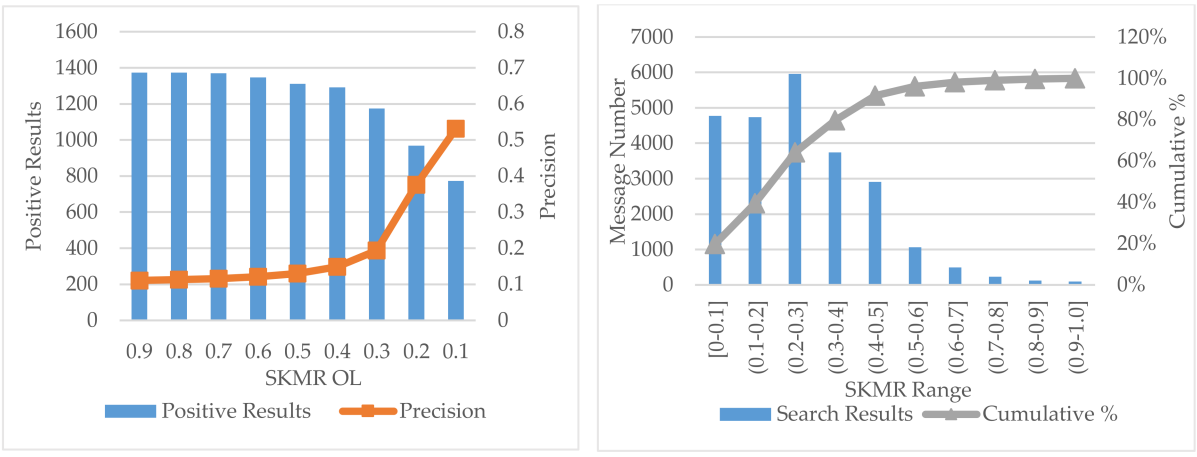


Figure 4. Influence of OROL values.

3.2.2. Single Keyword Message Ratio Overall Limit (SKMR OL)

Isometric values of SKMR OL from 0.9 to 0.1 were tested and are plotted independently from other filters in Figure 5a. Both positive results and precisions do not change considerably before the limit declines to 0.4. This result coincides with the distribution of search results in Figure 5b, as 80% of messages were aggregated in the range of [0, 0.4]. Therefore, SKMR OL is only effective in a specific range.



(**a**) (**b**)

Figure 5. (**a**) Effects of SKMR OL values; (**b**) Search Results distribution on different SKMR values.

3.2.3. Partial Limits

In Figure 6a, ORPL shows an unstable influence on precision, but a relatively regular effect for values under 0.3. In consideration of the distribution of SKMR values in Figure 5b, the numerical range of SKMR PL begins at 0.5, where 1.1 means no filter. Precision is enhanced when the limit value decreases, but the tendency reverses with small setbacks if SKMR PL is less than 0.2.

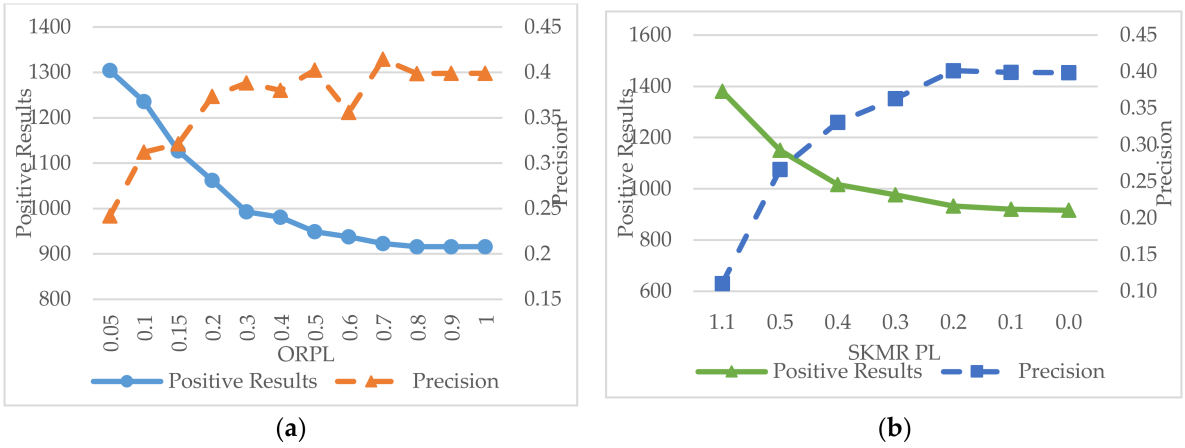


Figure 6. Performance of partial limits. (**a**) Influence of ORPL; (**b**) influence of SKMR PL.

In comparison with overall limits, partial limits performed with unsteadiness and limitations on precision. Positive effects of the partial limits sustain only in a local range. Furthermore, the maximum precision of overall limits is achieved at above 0.5, while that of partial limits swings around 0.4.

3.2.4. Combinations of the Filters

Since the main objective of the experiments was to find a set of filters optimizing the trade-off, the possibility of a better capability of a conjunct filter was explored after the individual impact of the filters was verified. The main principle is to select the option with the highest precision with a similar level of positive results.

The trade-offs of overall limits and their combinations are plotted in Figure 7, where the relation between positive results and precisions is revealed. All four groups of data manifest an obvious linearity, proved by their correlation coefficients smaller than −0.97. Apparently, the trendline that surpasses others will be the prime solution, which is the OROL.

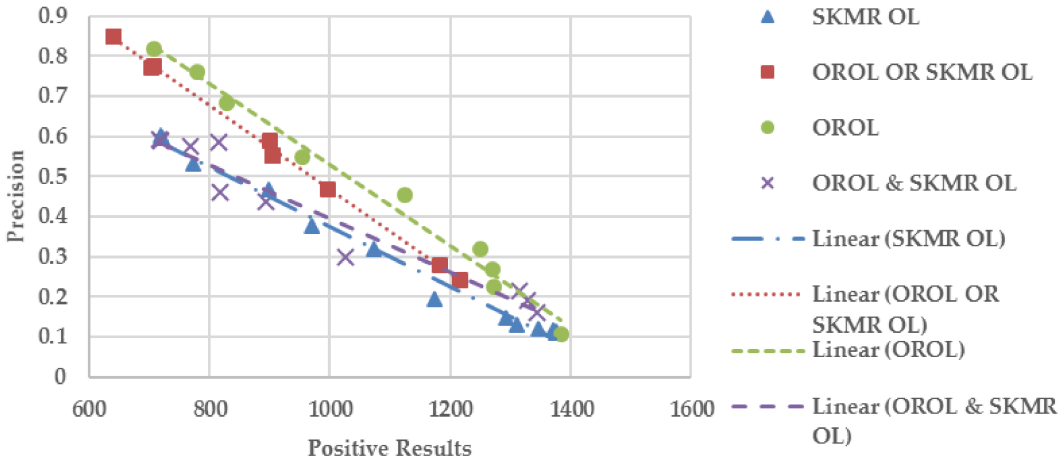


Figure 7. Trade-offs of overall limits.

Subsequent to the evaluation of overall limits, the optimum solution of partial limits should be similarly determined. However, the instabilities of their intersection or union both aggravate on the basis of individual limits. On the other hand, this randomness is partially alleviated when connected to OROL. Single partial limits have been stabilized with participation of OROL as in Figure 8. Additionally, their effective ranges have been widely expanded, and maximum precisions have

substantially improved. The same effects can be found for the intersection or union of partial limits. This enables the combinations of OROL and partial limits as possible options for an optimum solution.

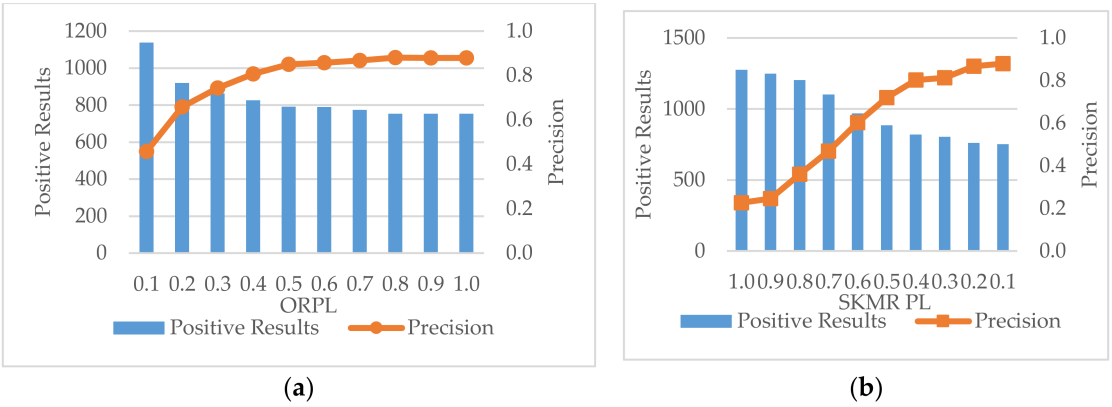


Figure 8. Partial limits combined with OROL. (**a**) Effect of ORPL where OROL = 0.0. (**b**) Effect of SKMR PL where OROL = 0.0.

As the precedent analysis has excluded some combinations, the remaining five options are OROL, OROL with ORPL, OROL with SKMR PL, OROL with ORPL & SKMR PL, and OROL with ORPL/SKMR PL. A scatter diagram, similar to Figure 6, was constructed and the area of high precision is more focused on in this figure, as it is the ultimate output precision of the filters. In Figure 9, the points of all combinations, not including OROL, almost overlap, which signifies small gaps. Therefore, the other four options combined with partial limits are more productive than the single OROL filter. This exception of OROL implies that partial limits perform more subtly than overall limits, while the latter contributes to moderating the labile capability of the former.

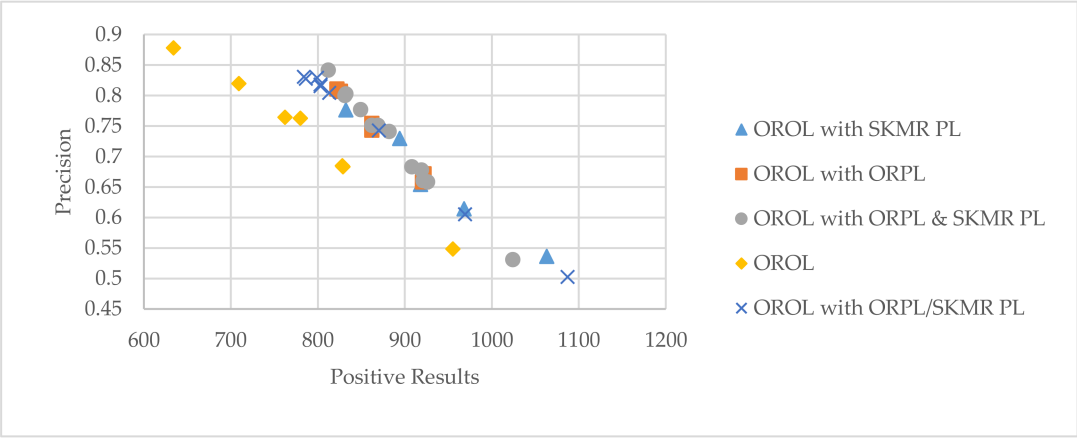


Figure 9. Trade-offs of overall limits.

3.2.5. Performance of Query Expansion

The search capacity of expanded query was contrasted with search results of other search strategies in Table 1. Among the tested filter pairs, the highest precision approximates to 0.9 with about a 30% reduction in positive results compared to the maximum above 0.5 precision. Test 3 represents a search by static keywords with no updates throughout, and it could act as a baseline to weigh coverage increase and precision reduction. Despite the minor decline in precision for Test 1, 34% more positive results were obtained. At an acceptable precision of 0.5 in Test 2, positive results almost doubled. The results support the effectiveness of the query expansion method.

Table 1. Performance of query expansion.

Test	OROL	SKMR PL	And/Or	ORPL	Positive Results	Precision	Description
1	0.05	0.3	Or	0.3	752	0.898	Maximum precision of the tested pairs
2	0.02	0.7	And	0.3	1090	0.507	Maximum positive results above 0.5 precision
3	-	-	-	-	561	0.996	Search by 'typhoon', 'hato' & 'pakhar'

Moreover, 2000 messages were randomly sampled from the original dataset of 27,665 tweets to assess the coverage of the method, and 102 messages were manually labelled as positive. These positive messages were compared to the search results with the parameters of Test 2 in Table 1. Among the 102 messages, 86 were classified as positive by the method, while the remaining 16 were neglected. Therefore, the estimated coverage is 86/102 = 84.3%.

3.3. Query Evolution

Except for the numerical evaluation by the precision and amount of positive results, it is crucial as well to examine the query evolution to ensure that the search process has not strayed away from the topic. The example of query evolution displayed below was randomly selected from tested combinations with the parameters in Table 2.

Table 2. Parameters of the search.

Type	Period	CKN	OROL	SKMR PL	And/Or	ORPL	Positive Results	Precision
Tested	12–60 h	50–250	−1, 0–0.35	0.1–0.9	And, Or	0.1–1	N/A	N/A
Selected	24 h	250	0.05	0.2	And	0.4	812	0.841

Table 3 manifests the query development in the first three rounds of the search. The total message number (TMN) represents the amount of data to be searched in the time span of the current round, and the search results are the quantities of positive messages. Only keywords of which search results account for more than 5% of the total are listed for succinctness, except Round 1 where the keywords are initial keywords stemming from news articles. The top three keywords retrieved only 16 messages in Round 1, indicating few discussions on typhoons in this period. In Rounds 2 and 3, the numbers of search results soared along with the increase in the TMN, and many symbolic keywords related to typhoons occur, such as 'signals'. In Round 2, 'signals' stands for the early warning stage of the disaster and the discussion focus shifts to typhoon impact in Round 3 by the rise of supply shortage keywords 'water', 'supply', and 'power' along with other keywords describing the disastrous situation, such as 'damage', 'affected', and 'hit'. Among the search results, prayers and safety checks prompted the usage of 'safe'. It should be noted that 'pigeon' in Round 2 is the literal meaning of the Japanese word 'hato', proving that the search strategy is capable of discovering new keywords in the local language context. In addition, 'Hong Kong' on the list was derived from the proximity in geographic locations and suffering from the same disaster with Macau.

Table 3. Query evolution of Rounds 1–3.

Round 1		**Round 2**		**Round 3**	
TMN [1]	SR [2]	TMN	SR	TMN	SR
2904	16	3160	265	3503	197
Keyword	**PCT [3]**	**Keyword**	**PCT**	**Keyword**	**PCT**
typhoon	75.00%	typhoon	63.02%	typhoon	50.76%
tropical	18.75%	hato	16.98%	kong	6.60%
hato	18.75%	hong	7.17%	hong	6.60%
signal	0%	kong	7.17%	water	15.74%
smg	0%	pigeon	8.30%	work	8.63%
meteorological	0%	signals	27.92%	power	6.60%
hoisted	0%			safe	6.09%
government	0%			supply	8.12%
bureau	0%			damage	9.64%
weather	0%			affected	9.64%
trajectory	0%			dead	5.58%
TdM [4]	0%			hit	8.63%

[1] total message number (TMN); [2] search results; [3] percentage of messages that the term is presented in the search results; [4] Teledifusão de Macau, a local television corporation at Macau.

Query evolution for the next three rounds is displayed in Table 4, where the TMN and the search results are both on the decline, suggesting the gradual vanishing of typhoon Hato. In Rounds 4 and 5, apart from the continuous supply problems, post-disaster situation descriptions and recovery work became the mainstream, while the word 'people' primarily denotes the People's Liberation Army of China who contributed to restoration after the calamity. The residents of Macau were actively aiding each other as implied by 'volunteer', 'group', and 'support'. However, denunciation on the government authorities for their deficient response surges on Twitter as manifested by the word 'government' in Round 4, and 'due' in Round 5 results from the phrasing of 'due to the influence of typhoon' in a great quantity of notices of closing down of shops and schools. In Round 6, recovery keywords mix with the early warning symbols 'signal', 'safety', and 'bureau', which is a fragment of Macau Meteorological Bureau, for the approaching typhoon Pakhar. The occurrence of 'pakhar' and 'paka', the transliteration of 'pakhar', is evidence that the search process remains on track and can detect variations in the discussion. Moreover, 'paka', resembling 'pigeon' in Round 2, is one of the diverse expressions of the same incident in different language environments.

In Table 5 for Rounds 7–9, the TMN falls back to the level of about 2800, and the search results likewise receded to the status before the typhoon swept the area. One thought-provoking fact is the variation in the search results proportion of the keyword 'typhoon'. In the period when the typhoon impacted Macau, it progressively dwindles but ascends with the obsolescence of the topic typhoon on Twitter. This may be owing to the participation of more discussants bringing in more text materials. On the other hand, this feature can be utilized in detecting the outburst of a topic on social media. In addition to keywords in the tables, some place names in Macau occur as query keywords, such as 'Taipa', 'Lisboa', 'Tongli', and 'Coloane'. Emergence of these locations could be signals of severe damage in these areas.

Table 4. Query evolution of Rounds 4–6.

Round 4		**Round 5**		**Round 6**	
TMN	SR	TMN	SR	TMN	SR
3225	154	2882	131	2816	110
Keyword	**PCT**	**Keyword**	**PCT**	**Keyword**	**PCT**
typhoon	44.81%	typhoon	38.93%	typhoon	53.64%
water	14.94%	people	16.79%	people	11.82%
work	9.09%	water	13.74%	work	9.09%
electricity	9.09%	pigeon	9.16%	wind	9.09%
kong	7.79%	clean	7.63%	storm	7.27%
hong	7.79%	work	13.74%	clean	5.45%
people	26.62%	volunteer	12.21%	lot	5.45%
clean	11.04%	newspaper	9.16%	staff	5.45%
government	14.94%	china	5.34%	car	6.36%
back	5.84%	relief	5.34%	home	8.18%
pigeon	9.74%	garbage	6.11%	bureau	7.27%
live	6.49%	damage	12.21%	street	5.45%
street	7.79%	open	6.87%	open	6.36%
disaster	14.29%	cities	15.27%	signal	19.09%
pay	5.19%	support	9.16%	pakhar	5.45%
post	7.14%	hours	5.34%	safety	6.36%
roads	5.19%	due	6.11%	small	6.36%
residents	5.19%	storm	6.11%	paka	6.36%
stopped	5.19%	lot	5.34%		
service	7.79%	group	7.63%		
areas	5.19%	wind	8.40%		
return	6.49%	areas	5.34%		

Table 5. Query evolution of Rounds 7–9.

Round 7		**Round 8**		**Round 9**	
TMN	SR	TMN	SR	TMN	SR
2700	44	2891	26	2956	22
Keyword	**PCT**	**Keyword**	**PCT**	**Keyword**	**PCT**
typhoon	43.18%	typhoon	80.77%	typhoon	72.73%
pigeon	6.82%	disaster	7.69%	disaster	18.18%
clean	6.82%	damage	19.23%	group	9.09%
hato	34.09%	newspaper	15.38%	large	13.64%
open	6.82%	clean	7.69%	garbage	9.09%
disaster	15.91%	group	11.54%	clean	18.18%
damage	15.91%	shimbun	7.69%	week	9.09%
support	15.91%	affected	11.54%	hong	18.18%
city	6.82%	small	7.69%	kong	18.18%
village	9.09%	information	7.69%	dam	9.09%
region	6.82%	analyst	7.69%	hato	22.73%
		impact	7.69%	due	9.09%
		garbage	7.69%	water	18.18%

The query evolution testifies that the query expansion method can cope with discussion focus shift and adapt to local language environment by discovering new keywords. The second capacity is especially vital for global social media platforms. During the processing of messages in multiple languages, translation is inevitable but it also results in information loss such as 'paka' or 'pigeon', which normally will not be associated with typhoon in English contexts. However, by the query expansion method, they are both recognized as relevant in this study. Occasionally, some foreign terms out of translation errors occur in query but are soon sifted out because they acquire no search results.

In Table 6, we explored whether the query expansion method could handle the variation in keywords along with the development of the topic. The number in each cell is the term frequency of a keyword in each round. Yellow background indicates that the keyword is used for searching in that round. In most cases when a keyword is trending it will be included as a query keyword. However, the delay between the rise in its term frequency and becoming a query keyword is noticeable, especially for 'water' and 'power'. A shorter period can reduce the omission caused by the delay because it will shorten the delay, which is supported by the results of Figure 2. However, consequent potential accuracy decrease in TF-IDF weighting for lack of materials should be considered as well.

Table 6. Term frequencies and query evolution.

Keywords	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	Round 9
signal	1	69	2	3	7	15	2	0	0
paka	0	0	0	0	7	6	2	0	0
pakhar	0	0	0	1	2	6	5	0	0
water	13	39	96	74	54	19	10	16	20
power	19	100	42	23	10	11	14	15	16
clean	0	3	18	19	10	13	9	3	3

3.4. Error Analysis

Although precision can be controlled by manipulating initial parameters and filters, it is also of significance to inspect the errors that are essentially irrelevant messages but classified otherwise in the search process to build more sophisticated filters. The composition of errors in the search results was analyzed to investigate the characteristics of false positives, which are classified into six categories below. User names are removed from the examples for privacy.

(1) Common words. In spite of their frequent application under circumstances of typhoon disasters, common words are generally used in other occasions. For instance, the keyword "work" continually occurs in descriptions of post-disaster restoration, but it as well introduces enormous false positives.

Example 1. *Macau gaming worker pay up 4.6 Percentage: govt data.*

(2) Semantic ambiguity. Messages in this category are likely to depict typhoons especially in the context of disaster; however, no decisive information is embodied in the text.

Example 2. *To Chen Mingjin Team: Macau is very lucky because of you!*

(3) Reference to other incidents. In this category, keywords depict incidents associated with typhoons, but explicit geographic or incident names oriented to other disasters are mentioned in the message.

Example 3. *Irma storm forms and threatens to become a hurricane heading for the Caribbean.*

(4) Words or word stems with multiple meanings. This category of errors often manifests as a coincident stem of two disparate words, such as a stem "car" for "careful" and "car", and the latter irrelevant word can generate numerous errors. Another case is the miscellaneous meanings of one word stem. For example, for a word stem "sign", it usually appears in early warnings as the word "signal". However, errors are produced by other words deriving from "sign" such as "sign" itself or "signature".

Example 4. *I added a video to a @YouTube playlist https://t.co/ToiH820cn9 Allegedly Learning to Drive, a Pickup Car Crashed into the Sea.*

Example 5. *Can you confirm I signed up for the haunted 5k and not 10K? Thanks!*

(5) Translation errors. Although it is of rare occurrence, defects exist in the translator particularly for proper nouns. In the example below, "天鴿" in Chinese represents the typhoon Hato, but occasionally it is inexplicably translated, even deviating from its literal meaning "heavenly pigeon". Even though the translator is frequently overwhelmed by the multilingual context and colloquial expressions on Twitter, errors in search results caused by translation is in the minority because translation errors are sporadic and cannot constitute a prevalent pattern in the whole data stream. Therefore, they can be easily eliminated via TF-IDF weighting and the filters.

Example 6. *Original: 天鴿一役揭示澳門行政改革的問題: 論盡媒體 AllAboutMacau Media https://t.co/dOEfcdLOAn. Translation: The Dayfoot Campaign Reveals Administrative Reform in Macao: On Media AllAboutMacau Media https://t.co/dOEfcdLOAn*

(6) Irrelevant keywords. Most keywords of weak relevance are filtered out by OROL, whereas the omitted minority turns out to be a source of errors.

Example 7. *A truck was driven into a store next to Loose. Officer said everyone is ok. Don't drink & drive.*

After labelling 500 random error messages, the result indicates that the proportion of common words outclass any other sources at the level of 88%. Multiple meanings account for 8.2%, while other incidents account for 2.4%. The other sources add up to a minority of less than 2%. This can be an area of focus in the design of more sensitive filters.

4. Conclusions

The main aim of this research was to develop a novel information retrieval method for the aggregation of geographic messages from social media by improving the TF-IDF weighting strategy. The proposed method was able to optimize the trade-off between precision and coverage of geographical messages by factoring in the influences of the keyword number and refresh cycle in the query process. First, the composition of short texts that are assumed relevant to a unified topic enhances the results of TF-IDF weighting, and the effectiveness of hypotheses and filters has been supported by experiments. Therefore, they can be utilized as independent classifiers for words or texts in various occasions such as text mining. Second, experiments indicate that a shorter Period and a larger Candidate Keyword Number can increase the coverage, although a set of increased query keywords leads to inefficiency owing to the overlap of their search results. Third, inspection of the query evolution demonstrates the abilities to remain pertinent to the topic, adapt to variation and local language contexts, and eliminate translation errors. The proposed approach manifests remarkable flexibility for social media streams. Finally, the trade-off between precision and coverage was optimized by testing different filters, and the query precision can be secured and regulated with effective filters. These filters ensure a reliable search process and show potential that the precision and coverage can be controlled for specific purposes, although more work on parameter value assignments is needed.

In conclusion, the proposed search method has accomplished the research objectives, and its effectiveness with controllable precision, more comprehensive data, and flexibility has been indicated. It also increases the credibility of subsequent spatial analysis. For practical applications, this method can be universally leveraged in information retrieval of social media messages or other text streams on various topics or incidents. Its capability of acquiring more relevant messages is particularly important in disaster management where a single important message can play a vital role in decision-making.

Further research is still required for the integrity of the approach. More datasets should be tested for the generalized validity of the filters and their proper numerical ranges. It should be noted that error analysis has provided some clues for improvement. For example, the less aggressive Porter stemmer

instead of the Lancaster algorithm can moderate the problem of word stems with multiple meanings, and appropriate coefficients can be assigned to common words when evaluating co-occurrence with other keywords to reduce their impact on filtering.

Acknowledgments: This research is sponsored by China Scholarship Council (CSC).

Author Contributions: Zi Chen conceived and designed the experiments, performed the experiments, analyzed the data, and wrote the paper; Samsung Lim offered advice in the experiments and modification to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodchild, M.F. Citizens as sensors: The world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
2. Sui, D.; Goodchild, M. The convergence of GIS and social media: Challenges for GIScience. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1737–1748. [CrossRef]
3. Maresh-Fuehrer, M.M.; Smith, R. Social media mapping innovations for crisis prevention, response, and evaluation. *Comput. Hum. Behav.* **2016**, *54*. [CrossRef]
4. Simon, T.; Goldberg, A.; Adini, B. Socializing in emergencies—A review of the use of social media in emergency situations. *Int. J. Inf. Manag.* **2015**, *35*, 609–619. [CrossRef]
5. Towards Real-time Emergency Response using Crowd Supported Analysis of Social Media. Available online: https://www.researchgate.net/publication/228975334_Towards_Real-time_Emergency_Response_using_Crowd_Supported_Analysis_of_Social_Media (accessed on 20 January 2018).
6. Deng, Q.; Liu, Y.; Zhang, H.; Deng, X.; Ma, Y. A new crowdsourcing model to assess disaster using microblog data in typhoon Haiyan. *Nat. Hazards* **2016**, *84*, 1241–1256. [CrossRef]
7. Yin, J.; Lampert, A.; Cameron, M.; Robinson, B.; Power, R. Using Social Media to Enhance Emergency Situation Awareness. *IEEE Intell. Syst.* **2012**, *27*, 52–59. [CrossRef]
8. Chowdhury, R.; Chowdhury, S.R.; Castillo, C. Tweet4act: Using Incident-Specific Profiles for Classifying Crisis-Related Messages. In Proceedings of the 10th International ISCRAM Conference; ISCRAM: Baden, Germany, 2013; pp. 834–839.
9. Vieweg, S.; Castillo, C.; Imran, M. Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014. Proceedings*; Aiello, L.M., McFarland, D., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 444–461. ISBN 978-3-319-13734-6.
10. Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; Rogstadius, J. Coordinating human and machine intelligence to classify microblog communications in crises. In *ISCRAM 2014 Conference Proceedings—11th International Conference on Information Systems for Crisis Response and Management*; ISCRAM: University Park, PA, USA, 2014; pp. 712–721.
11. Vechtomova, O.; Wang, Y. A study of the effect of term proximity on query expansion. *J. Inf. Sci.* **2006**, *32*, 324–333. [CrossRef]
12. Cui, H.; Wen, J.-R.; Nie, J.-Y.; Ma, W.-Y. Probabilistic query expansion using query logs. In Proceedings of the Eleventh International Conference on World Wide Web—WWW'02, Honolulu, HI, USA, 7–11 May 2002; ACM Press: New York, NY, USA, 2002; p. 325.
13. Rivas, A.R.; Iglesias, E.L.; Borrajo, L. Study of query expansion techniques and their application in the biomedical information retrieval. *Sci. World J.* **2014**, *2014*, 132158. [CrossRef] [PubMed]
14. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008; ISBN 0521865719, 9780521865715.
15. Harris Smith, S.; Bennett, K.J.; Livinski, A.A. Evolution of a Search: The Use of Dynamic Twitter Searches During Superstorm Sandy. *PLoS Curr.* **2014**. [CrossRef] [PubMed]
16. Lin, C.X.; Zhao, B.; Mei, Q.; Han, J. PET: A Statistical Model for Popular Events Tracking in Social Communities. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 25–28 July 2010; ACM: New York, NY, USA, 2010; pp. 929–938.

17. Massoudi, K.; Tsagkias, M.; de Rijke, M.; Weerkamp, W. Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. In *Advances in Information Retrieval*; Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 362–367.
18. Zhao, L.; Chen, F.; Lu, C.T.; Ramakrishnan, N. Dynamic theme tracking in Twitter. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 29 October–1 November 2015; pp. 561–570.
19. Mei, Q.; Zhai, C. Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining. In Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA, 21–24 August 2005; ACM: New York, NY, USA, 2005; pp. 198–207.
20. Bandyopadhyay, A.; Ghosh, K.; Majumder, P.; Mitra, M. Query expansion for microblog retrieval. *Int. J. Web Sci.* **2012**, *1*, 368. [CrossRef]