Estimating Document Focus Time

Adam Jatowt, Ching-Man Au Yeng, Katsumi Tanaka

October 29, 2016

Overview

1. Introduction
2. Estimation of Focus Time
3. Experimental Setting
4. Results
5. Conclusion

Introduction

- **What** is Focus Time of a Document?

Introduction

- **What** is Focus Time of a Document?
  - ⋆ Time period to which document **content's** refers, i.e, the relation of document **content** with a time period

Introduction

- **What** is Focus Time of a Document?
  - ⋆ Time period to which document **content's** refers, i.e, the relation of document **content** with a time period
  - ⋆ **Not** the document creation time or last modified time stamp

Introduction

- **What** is Focus Time of a Document?
  - ⋆ Time period to which document **content's** refers, i.e, the relation of document **content** with a time period
  - ⋆ **Not** the document creation time or last modified time stamp
- **Why** do we need it?
  - ⋆ Search query having temporal characters
  - ⋆ Search engine need to match time scope of query with the document

Introduction

- **What** is Focus Time of a Document?
  - ⋆ Time period to which document **content's** refers, i.e, the relation of document **content** with a time period
  - ⋆ **Not** the document creation time or last modified time stamp
- **Why** do we need it?
  - ⋆ Search query having temporal characters
  - ⋆ Search engine need to match time scope of query with the document
  - ⋆ Solution1: return the documents which explicitly contains date with in the time scope.

Introduction

- **Why** this is not optimal?

Introduction

- **Why** this is not optimal?
  - ⋆ Document does not have any or have few temporal expressions

Introduction

- **Why** this is not optimal?
  - ⋆ Document does not have any or have few temporal expressions
- Solution2: Consider Creation time.
  - ⋆ Temporal information is neglected if only creation time stamp is used
    - ⋆ Example a news report on an event that is already dated.

Introduction

- **Why** this is not optimal?
  - ⋆ Document does not have any or have few temporal expressions
- Solution2: Consider Creation time.
  - ⋆ Temporal information is neglected if only creation time stamp is used
    - ⋆ Example a news report on an event that is already dated.
- More over time plays a central role in several area like information extraction, topic detection, question answering,summarization etc.

Introduction

- **Why** it is difficult ?

Introduction

- **Why** it is difficult ?
  - In many cases there are very few or no temporal expression in the document
  - The document itself is atemporal eg. tutorial on linear algebra

Introduction

[1] Ack: Generic Method for calculating document focus time , Jatowt et al.

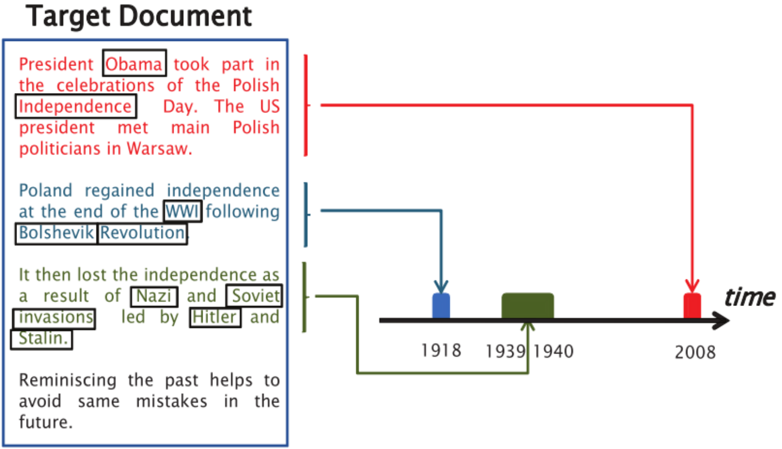
Introduction

- Consider a hypothetical Document

[1] Ack: Generic Method for calculating document focus time , Jatowt et al.

Introduction

- Consider a hypothetical Document



Target Document

President Obama took part in the celebrations of the Polish Independence Day. The US president met main Polish politicians in Warsaw.

Poland regained independence at the end of the WWI following Bolshevik Revolution.

It then lost the independence as a result of Nazi and Soviet invasions led by Hitler and Stalin.

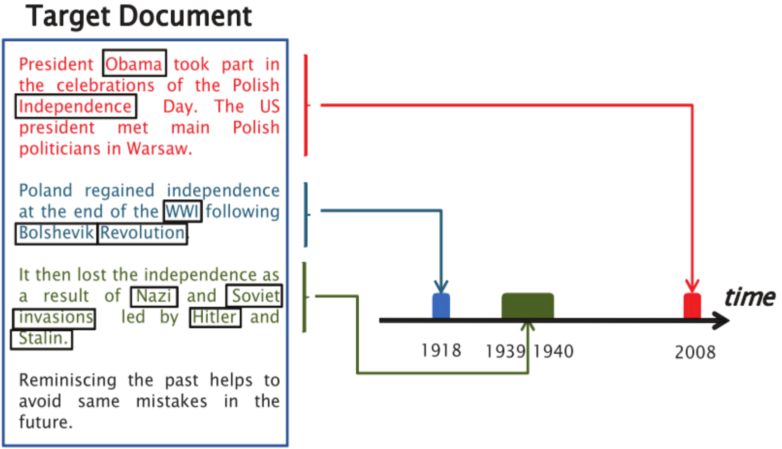
Reminiscing the past helps to avoid same mistakes in the future.

Fig. Mapping content of an example document onto timeline.

[1] Ack: Generic Method for calculating document focus time , Jatowt et al.

Introduction

- Consider a hypothetical Document



Target Document

President Obama took part in the celebrations of the Polish Independence Day. The US president met main Polish politicians in Warsaw.

Poland regained independence at the end of the WWI following Bolshevik Revolution.

It then lost the independence as a result of Nazi and Soviet invasions led by Hitler and Stalin.

Reminiscing the past helps to avoid same mistakes in the future.

Fig. Mapping content of an example document onto timeline.[1]

[1] Ack: Generic Method for calculating document focus time , Jatowt et al.

Introduction

- It contains mixture of sentences referring to past events
- It contains sentences which are atemporal

Introduction

- It contains mixture of sentences referring to past events
- It contains sentences which are atemporal
- But none of the sentence contains any explicit or implicit temporal expression

Introduction

- It contains mixture of sentences referring to past events
- It contains sentences which are atemporal
- But none of the sentence contains any explicit or implicit temporal expression
- As a human we can position its content onto time line (as indicated)using temporal clue word "Obama", "Nazi", "Soviet" and "Stalin"
- Moreover, we can some what infer that the document is mainly focusing on Polish Independence day

Introduction

- Now the important question to ask is ”Can the same process be done automatically?”

Introduction

- Now the important question to ask is "Can the same process be done automatically?"
- **Benefits**

Introduction

- Now the important question to ask is ”Can the same process be done automatically?”
- **Benefits**
  - ⋆ Categorizing documents by their temporal foci and mapping their content onto timeline.
  - ⋆ Improves the performance of search engine in handling temporal quires
  - ⋆ Helps in document understanding.
  - ⋆ Have important applications on document summerization, information extraction, question answering, and so on.
  - ⋆ Example creating a particular historical period, describing chronological context of the text, can be used as extension to existing ranking technique.

Background

- The concept of focus time is defined as.

Defination

A temporal document, d, has the **focus time** τ if its content refers to τ

Background

- The concept of focus time is defined as.

Defination

A temporal document, d, has the **focus time** τ if its content refers to τ

→ This means for example that the document describes events which occurred or person who lived in a given time period.

Background

- This corpus-based statistical method is composed of three steps

Background

- This corpus-based statistical method is composed of three steps
    1. Calculating the strength of *word-time* association based on statistical knowledge derived from external document collection
        - ⋆ "Nazi" and "Hitler" are strongly related to time period 1939 to 1945

Background

- This corpus-based statistical method is composed of three steps
  1. Calculating the strength of *word-time* association based on statistical knowledge derived from external document collection
     - ⋆ "Nazi" and "Hitler" are strongly related to time period 1939 to 1945
  2. Estimating temporal weights of words
     - ⋆ For selecting discriminating terms which should be most helpful in estimating focus time(eg., temporal clue word in the previous Figure)

Background

- This corpus-based statistical method is composed of three steps
  1. Calculating the strength of *word-time* association based on statistical knowledge derived from external document collection
     - ⋆ "Nazi" and "Hitler" are strongly related to time period 1939 to 1945
  2. Estimating temporal weights of words
     - ⋆ For selecting discriminating terms which should be most helpful in estimating focus time(eg., temporal clue word in the previous Figure)
  3. Calculating text focus time
     - ⋆ Final estimation is done by extrapolating term focus point to document focus time by set of combination methods, ie. finding synchronicity between different temporal pointers

Measuring *word-time* associations

- For this, we use an external knowledge base which contains reference to past associated with absolute dates
- Large dataset of news article on diverse topic is used as resource
- Construct weighted, undirected graph $G(V, E)$
  - ⋆ V denotes the set of vertices being the vocabulary of the news
  - ⋆ E is set of edges representing word co-occurrences

Measuring *word-time* associations

- For this, we use an external knowledge base which contains reference to past associated with absolute dates
- Large dataset of news article on diverse topic is used as resource
- Construct weighted, undirected graph $G(V, E)$
  - ⋆ V denotes the set of vertices being the vocabulary of the news
  - ⋆ E is set of edges representing word co-occurrences

Measuring *word-time* associations

- Edge is labeled with the weights calculated from Jaccard Coefficient

$$A_{dir}(w_i, w_j) = \frac{c(w_i, w_j)}{c(w_i) + c(w_j) - c(w_i, w_j)}$$

  - ⋆ $c(w_i, w_j)$ = count of sentence where w_i,w_j co-occur
  - ⋆ $c(w_i), c(w_j)$ = number of sentence containing words w_i,w_j

- Here the dates occurring in the news article are treated as words
- Thus a word w can be associated with an arbitrary time point t which indicates a particular year denoted as $A_{dir}(w_i, t)$

Measuring *word-time* associations

- Calculating word-time association this way results in sparse results due to relatively small number of dates as compared to number of words in the document
- Therefore we extend this approach by considering the words context, that is, other words that strongly co-occur with a given word
- This way we are also utilizing semantic similarity of words based on their context

Intuition

Word w is strongly associated with time point t if many other words that strongly co-occur with w are also strongly associated with t

Measuring *word-time* associations

- Formula for context based association ,$A_{con}(w_i, t)$ is as follows

$$A_{con}(w_i, t) = \frac{1}{|V|} \sum_{j=1}^{|V|} A_{dir}(w_i, w_j)^2 A_{dir}(w_j, t)$$

  - ⋆ $|V|$ is total number of vertices(words)
  - ⋆ Taking squares of the values of $A_{dir}(w_i, w_j)$ decreases the impact of terms which are weakly associated with the target word w_i

- Lastly, to normalize the scores of words with each time point we divide them by the geometric mean of the association scores of all words with this time point

Estimating Temporal Weight

- We need to give some measure for the word importance with respect to time
  - ⋆ Not all words will be equally useful to determine focus time
  - ⋆ A term such as "earthquake" or "war" would have higher score than that of "tree" or "sun"
  - ⋆ To effectively determine the usefulness of word in discriminating time

Estimating Temporal Weight

- We need to give some measure for the word importance with respect to time
  - ⋆ Not all words will be equally useful to determine focus time
  - ⋆ A term such as "earthquake" or "war" would have higher score than that of "tree" or "sun"
  - ⋆ To effectively determine the usefulness of word in discriminating time
- So we follow the following assumption after analyzing the word association with time

Assumption

A word has high discriminative capability for determining document focus time if it has strong association with only few time points and weak association with other time points

Estimating Temporal Weight

- Thus a word have high discriminating capability if it strongly points to one or only few time points
- To rank term according to their discriminating capability we compute *temporal entropy* over association score of word with all time points
  - ⋆ We normalize the association scores to obtain the probability distribution over time
  - ⋆ Given a word, divide its association score with particular time point by sum of word's association score with all time points
  - ⋆ That is the Probability that a word w is associated with the time point t_i, given by $P_w(t_i)$

Estimating Temporal Weight

- We define temporal entropy ϖ_w^{temE} as

$$E_w = \varpi_w^{temE} = -\sum_i P_w(t_i) \ln P_w(t_i)$$

  ⋆ $P_w(t_i)$ is probability the word w is associated with time t_i

Estimating Temporal Weight

- It favors words that have non-uniform probability distribution over association with time
  - Because word having uniform association with different years are clearly not useful

Estimating Temporal Weight

- It favors words that have non-uniform probability distribution over association with time
  - Because word having uniform association with different years are clearly not useful
- But! this measure does not consider the distance between peaks in word-time associations
  - ⋆ "earthquake" or "war" would have long distance between their peak in word-time probability distribution
- Relying on them may bring confusion and hinder the performance of focus time

Estimating Temporal Weight

- Thus it is better to find terms having strong association with few nearby years or only one year (e.g, "Einstein" or "Hurricane Katrina")

Estimating Temporal Weight

- Thus it is better to find terms having strong association with few nearby years or only one year (e.g, "Einstein" or "Hurricane Katrina")
- To reflect this, as second term measure, *temporal kurtosis* is introduced which favors the words having distribution with one high peak ("Kurtosis is a measure of tailedness of probability distribution ")

Estimating Temporal Weight

- Thus it is better to find terms having strong association with few nearby years or only one year (e.g, "Einstein" or "Hurricane Katrina")
- To reflect this, as second term measure, *temporal kurtosis* is introduced which favors the words having distribution with one high peak ("Kurtosis is a measure of tailedness of probability distribution ")

$$K_w = \varpi_w^{temK} = \frac{\sum_i (A(w, t_i) - \mu)^4}{N\sigma^2}$$

⋆ N = total number of time point

⋆ μ, σ = mean and standard of word-year association distribution

⋆ $A(w, t_i)$ = any association score as described previously

Calculating Document-Time Association

- Once we obtain the word-time association and compute their temporal weights, we proceed to find focus time of document
- The basic intuition for the calculating document-time association is

Intuition

The more words strongly associated with time point t are contained in a document d, the more it is likely that t belongs to the focus time of d

Calculating Document-Time Association

- This hypothesis is implemented by computing the weighted avgerage over scores representing the association of document terms with time

Calculating Document-Time Association

- This hypothesis is implemented by computing the weighted avgerage over scores representing the association of document terms with time
- In first method only unique word in the document is considered

$$S_U(d,t) = \frac{1}{|d|} \sum_{w \in d} \varpi_w^{tem} A(w,t)$$

  - ⋆ $S_U(d,t)$ is the score of a time point t when using all unique term in text document d
  - ⋆ $|d|$ is vocabulary size of the document
  - ⋆ This score is regarded as "goodness" of t belong to focus time of d

Calculating Document-Time Association

- Extension of this approach is by considering frequency of the term in the document
- That is given by the equation

$$S_{TF}(d,t) = \frac{1}{\sum_{w \in d} N(w,d)} \sum_{w \in d} \varpi_w^{tem} N(w,d) A(w,t)$$

  ⋆ N(w,d) is number of times word w has occurred in the document d

Calculating Document-Time Association

- Term frequency measure alone cannot distinguish representative keywords.
- Clue words (past entities) may sometime appears sparsely despite being crucial for estimating the relevant time period
- Thus we need additional weight-age that would give word importance in the text
  - ⋆ Use of **Text rank** is intuitive for considering importance of word in the text

Calculating Document-Time Association

- Text-rank is recursive calculation over word interconnection in the document
- The score is given by

$$S_{TR}(d,t) = \frac{1}{\sum_{w \in d} N(w,d)} \sum_{w \in d} \varpi_w^{imp} \varpi_w^{tem} N(w,d) A(w,t)$$

  ⋆ ϖ_w^{imp} is the importance of word calculated by text-rank

- This equation represents the document-time association scores based on both "importance" as well as "temporal weights"

Calculating Document-Time Association

- Note : all the association score introduced so far do not explicitly use temporal expression

Calculating Document-Time Association

- Note : all the association score introduced so far do not explicitly use temporal expression
- Temporal expression such as dates appearing in the document constitutes important signals of document focus time

Calculating Document-Time Association

- Note : all the association score introduced so far do not explicitly use temporal expression
- Temporal expression such as dates appearing in the document constitutes important signals of document focus time
- To utilize it we extract all the date from the document

Calculating Document-Time Association

- Note : all the association score introduced so far do not explicitly use temporal expression
- Temporal expression such as dates appearing in the document constitutes important signals of document focus time
- To utilize it we extract all the date from the document
- Apply Gaussian Kernel Density Estimate using extracted date
  - ⋆ It generates Gaussian distribution centered at extracted dates

Calculating Document-Time Association

- Note : all the association score introduced so far do not explicitly use temporal expression
- Temporal expression such as dates appearing in the document constitutes important signals of document focus time
- To utilize it we extract all the date from the document
- Apply Gaussian Kernel Density Estimate using extracted date
  - ⋆ It generates Gaussian distribution centered at extracted dates
- Mixture of such distribution is then considered as date-based document-time association. Denoted as $S_{DATE}(d, t)$
- Then we propose the extended document-time association score

$$S_{EXT}(d, t) = S_{DATE}(d, t) + S(d, t)$$

Computing Document Focus Time

- After computing document-time association score we can proceed to estimate the document focus time

Computing Document Focus Time

- After computing document-time association score we can proceed to estimate the document focus time
- As a simple implementation , we choose only single time point with highest association of focus time and denoted as $t_{foc}(d)$

Computing Document Focus Time

- After computing document-time association score we can proceed to estimate the document focus time
- As a simple implementation , we choose only single time point with highest association of focus time and denoted as $t_{foc}(d)$
- It is given by the equation

$$t_{foc}(d) = argmax_t S(d, t)$$

Preparing Knowledge Base

- Requires a sufficient amount of temporally grounded temporal reference for calculating the *word-time association*

Preparing Knowledge Base

- Requires a sufficient amount of temporally grounded temporal reference for calculating the *word-time association*
  - ⋆ News article published from 1990 to 2010 from Google News Archive.
  - ⋆ Used country names instead of arbitrary queries because most of the events takes place in particular countries (Germany, UK, Japan, France and Israel)
  - ⋆ By choosing countries there will be no restriction on the type of events
  - ⋆ Total of 535K news articles where fetched across the all countries(Germany:87K, UK:149K, France: 110K, Japan: 97K and Israel: 92K)

Preparing Knowledge Base

- Requires a sufficient amount of temporally grounded temporal reference for calculating the *word-time association*
  - ⋆ News article published from 1990 to 2010 from Google News Archive.
  - ⋆ Used country names instead of arbitrary queries because most of the events takes place in particular countries (Germany, UK, Japan, France and Israel)
  - ⋆ By choosing countries there will be no restriction on the type of events
  - ⋆ Total of 535K news articles where fetched across the all countries(Germany:87K, UK:149K, France: 110K, Japan: 97K and Israel: 92K)
- This knowledge base is used for collecting statistical information for estimating the focus times of test documents.

Document Datasets for Testing

- Once the *word-time association* is established we need data to TEST the approach
- The data for this purpose is taken from 3 sources
  1. Wikipedia
  2. Book Dataset
  3. Web Dataset

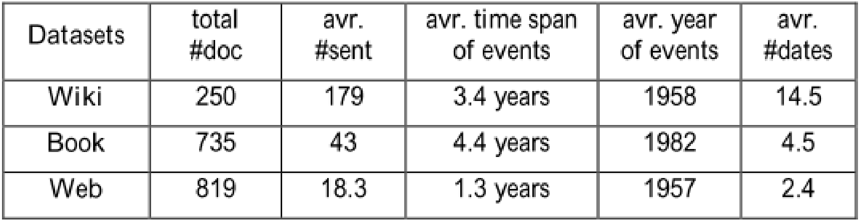
Document Datasets for Testing

- Wikipedia Dataset Group
  - ⋆ Total 250 article devoted to major historical events related to selected countries where fetched(50 for each country)
  - ⋆ The event in these article occurred with in time frame of 1900 to 2013
  - ⋆ Event were of different types including major wars, battles, treaties, strikes, elections, and any other
  - ⋆ Wikipedia provide good source for evaluation purpose as they contain precise metadata in form of start and end dates.

Document Datasets for Testing

- Book Dataset Group
  - ⋆ Used two history related books: "Timeline of World History" and "Timelines of History"
  - ⋆ It covers key events of last centuries in short paragraph that are ordered chronologically
  - ⋆ Had to extract sentences containing the name of any of the five selected countries
- Web Dataset Group
  - ⋆ Created using popular history-focused websites: "History Orb", "History World", "BBC Timelines" and "Infoplease"
  - ⋆ Regarded each paragraph as separate document and assigned dates corresponding to it either from title or manually

Document Datasets for Testing



Datasets	total #doc	avr. #sent	avr. time span of events	avr. year of events	avr. #dates
Wiki	250	179	3.4 years	1958	14.5
Book	735	43	4.4 years	1982	4.5
Web	819	18.3	1.3 years	1957	2.4

Figure: Dataset statistics (aggregated over all countries)

[2]

[2] Estimating Document focus time, Jatowt et al.

Baselines

- For comparing the effectiveness of the approach two baselines are used

Baselines

- For comparing the effectiveness of the approach two baselines are used
  1. Random Baseline
     - ⋆ : Randomly estimates focus time as a random year within the set time period
  2. Date-based baseline
     - ⋆ Utilizes only absolute dates occurring in text disregarding the rest of the context of the text.
     - ⋆ This is similar as $S_{Date}(d, t)$

Evaluation Measure

- Error of estimating the focus time

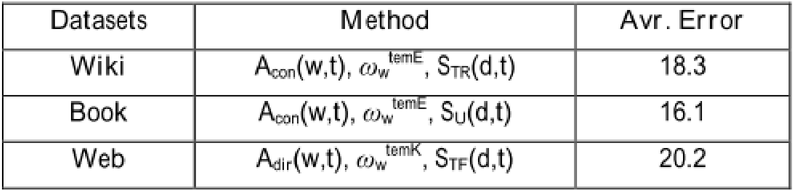
$$e(t_{foc}) = \begin{cases} \min\{|t_b - t_{foc}|, |t_{foc} - t_e|\} & \text{if } t_{foc} \notin [t_b, t_e] \\ 0 & \text{if } \textit{otherwise} \end{cases}$$

⋆ Where the time period $[t_b, t_e]$ is the actual focus time given by ground truth data

⋆ The higher the error means that t_{foc} is farther from ground truth

Results

- First, various combination of the proposed approaches which do not rely on the presence of temporal expression has been tested.
- Since the number of combination are high, only the best result per dataset has been shown



Datasets	Method	Avr. Error
Wiki	$A_{con}(w,t)$, ω_w^{temE}, $S_{TR}(d,t)$	18.3
Book	$A_{con}(w,t)$, ω_w^{temE}, $S_{U}(d,t)$	16.1
Web	$A_{dir}(w,t)$, ω_w^{temK}, $S_{TF}(d,t)$	20.2

Figure: Average error of the best method in focus time estimation

[3]

[3] Estimating Document focus time, Jatowt et al.

Results

- These results are quite satisfactory considering relatively long length of the time span (over 110 years)
- The average length of the described events are short (3.4, 4.4, 1.3 years in the Wiki, Book and Web datasets, respectively)

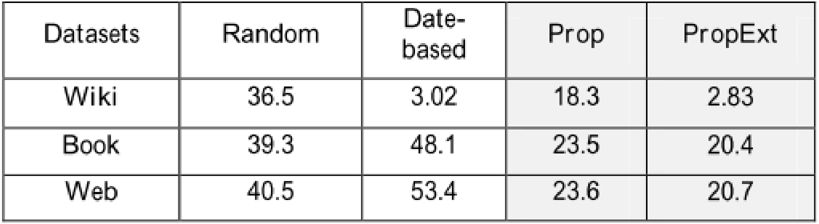
Results

- These results are quite satisfactory considering relatively long length of the time span (over 110 years)
- The average length of the described events are short (3.4, 4.4, 1.3 years in the Wiki, Book and Web datasets, respectively)
- **Observations**
  - $A_{con}(w, t)$ works well in most of the cases which implies that using the year associations of words that are strongly co-occurring with the target word seems to help better estimate the focus time
  - Text-rank method is useful when the document length is large
  - In case of short length of the document term-frequency or unique term performs better

Results

- Now comparison of the proposed approaches against the baselines has to be performed
- For the comparison the chosen method is $(A_{con}(w,t), \varpi_w^{tem}, S_{TR}(d,t))$ because it performed consistently well on different datasets
- we use extended version that incorporates dates occurring in text $(PropExt)$ and the one that simply use the only terms $(Prop)$

Results



Datasets	Random	Date-based	Prop	PropExt
Wiki	36.5	3.02	18.3	2.83
Book	39.3	48.1	23.5	20.4
Web	40.5	53.4	23.6	20.7

Figure: Comparison with baselines

[4]

[4] Estimating Document focus time, Jatowt et al.

Results

- **Observations:**
  - Prop achieves better results than the baselines for all the datasets (except date-based on Wiki data-set)
    - Wikipedia articles contain relatively many dates; hence, the straightforward date-based baseline performs better
  - The Book and Web datasets, the date-based baseline performs very poorly
    - many texts about past may not contain any temporal expressions or may contain only few of them
  - The PropExt performs best
    - Because it consider dates explicitly occurring in the text along with the terms.
  - The date-base approach and the PropExt performs almost equally on Wikipedia dataset.
    - Implies that when sufficiently many dates are present in text, as is in the case of the Wikipedia documents, the improvement over the date-based approach may not be significant.

Conclusion

- Applications
  - Document focus time could be considered as a complement to semantic representation of documents.
  - Can be used to calculate temporal similarity between two documents in parallel to their content similarity.
  - Temporal representation of a query could be matched to the one of document.
  - Improving ordering of sentences in automatically created document summaries
  - Adding single dates to an inverted index should have relatively small effect on the index size,while computing the temporal similarity between a query and a document would be relatively fast.

Conclusion

- Described the concept of document focus time and provide a range of methods for its estimation
- Approach uses corpus statistics, especially it uses absolute references to past years in news articles
- This method also works for documents which do not contain any temporal expressions
- The experimental evaluation indicates that the proposed methods provide satisfactory results over diverse sets of documents

References

Estimating Document Focus Time (CIKM 2013), Adam Jatowt, Ching-Man Au Yeng, Katsumi Tanaka
Generic method for detecting focus time of a documents.(Journal:IPM 2015), Adam Jatowt, Ching-Man Au Yeng, Katsumi Tanaka.

Thanks