Evaluating Significance of Historical Entities Based on Tempo-Spatial Impacts Analysis Using Wikipedia Link Structure

Yuku Takahashi
Graduate School of Infomatics
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto, Kyoto, Japan
ytakahas@dl.kuis.kyoto-u.ac.jp

Hiroaki Ohshima
Graduate School of Infomatics
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto, Kyoto, Japan
ohshima@dl.kuis.kyoto-u.ac.jp

Mitsuo Yamamoto
Denso IT Laboratory
Shibuya Cross Tower 28th
Floof 2-15-1
Shibuya, Shibuya-ku, Tokyo,
Japan
miyamamoto@d-itlab.co.jp

Hirotoshi Iwasaki
Denso IT Laboratory
Shibuya Cross Tower 28th
Floof 2-15-1
Shibuya, Shibuya-ku, Tokyo,
Japan
hiwasaki@d-itlab.co.jp

Satoshi Oyama
Graduate School of
Information Science and
Technology
Hokkaido University
Kita 14, Nishi 9, Kita-ku,
Sapporo, Hokkaido, Japan
oyama@ist.hokudai.ac.jp

Katsumi Tanaka
Graduate School of Infomatics
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto, Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT

We propose a method to evaluate the significance of historical entities (people, events, and so on.). Here, the significance of a historical entity means how it affected other historical entities. Our proposed method first calculates the tempo-spacial impact of historical entities. The impact of a historical entity varies according to time and location. Historical entities are collected from Wikipedia. We assume that a Wikipedia link between historical entities represents an impact propagation. That is, when an entity has a link to another entity, we regard the former is influenced by the latter. Historical entities in Wikipedia usually have the date and location of their occurrence. Our proposed iteration algorithm propagates such initial tempo-spacial information through links in the similar manner as PageRank, so the tempo-spacial impact scores of all the historical entities can be calculated. We assume that a historical entity is significant if it influences many other entities that are far from it temporally or geographically. We demonstrate a prototype system and show the results of experiments that prove the effectiveness of our method.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

Wikipedia structure analysis, Historical entities, PageRank, Historical entity importance

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.

1. INTRODUCTION

"What event is the most significant on the history of mankind?" "Which is greater, *Albert Einstein* or *Isaac Newton*?" This kind of historical questions are quite interesting but difficult to answer. This is because the questions are subjective and their answers depend on viewpoints and answerers[1].

We explore ways to evaluate the *significance* of historical entities in a more objective manner from the viewpoint of the degree of *tempo-spatial impact*. Here, historical entities (as defined in Section 3.1) are people, organizations, events, places, building structures, and so on. Our proposed method is based on the following hypothesis: the more significant a historical entity is, the more entities must be affected by it. Furthermore, the influence of a significant historical entity is widespread both temporally and geographically. That is, it affects many entities that are in different time periods or that are located in different places.

Based on this hypothesis, we first propose a method for calculating the tempo-spacial *impact* of historical entities. The impact of a historical entity varies according to time and location. Basically, the influence of a historical entity

[1]For example, *Isaac Newton* is considered to have had a greater impact on both science and humankind than *Albert Einstein*, according to the results of two Royal Societies in London polls announced 23 November 2005. This debate was part of the Einstein Year celebrations.

becomes weaker when time and location are farther from the ones of another entity. For example, *Napoleon* had strong influence in Europe in his era. However, he had almost no influence in Japan in that era, and he seems to have a little influence in Europe currently. We focus on the reference relationships among historical entities. Each article in Wikipedia is treated as a single historical entity because it usually covers differing historical entities. Each link between two Wikipedia articles is considered as a reference relationship. Impact is propagated through the Wikipedia links. That is, when an entity has a link to another one, we regard the former is influenced by the latter. Some historical entities in Wikipedia have the date and location of their occurrence. Our iteration algorithm propagates such initial tempo-spacial information through links in the similar manner as PageRank, so the tempo-spacial impact scores of all the historical entities can be calculated.

The tempo-spacial impact scores of an entity can be restricted by time or location such as "*Newton* in 1900" or "*Napoleon* on England." Applications that show temporal or geographical change of impact scores of an entity on a graph or on a map are available. We think that this type of service can be extended to not only how historical entities gave an impact on the present day but also to help us to understand what future impact might be. For example, it could help to provide people with estimation on how the visited places by them influenced their home places.

The representative time and location of each historical entity can be estimated from the impact. We assume that a historical entity is significant if it influences many other entities that are far from it temporally or geographically. We propose a method for evaluating the significance of historical entities according to this assumption.

We implemented a system and compared the proposed method with several baseline methods. The result of experiments demonstrates the effectiveness of our proposed method.

2. RELATED WORK

In this research, we use Wikipedia link structure for evaluating significance of historical entities.

PageRank [1] is a method to compute significance of web pages using web link structure. In PageRank, a hyperlink from a web page to the target web page is regarded as a mind of support. The assumption underlying PageRank algorithm is that the more a web page is linked, the more reliable that web page is, and that the web page linked by many reliable web pages is even more reliable. The authors propose an iterative way to calculate PageRank. Further extensions to the concept of biased PageRank were proposed, such as TrustRank [20], and topic-sensitive PageRank [6]. TrustRank by Gyöngyi et al. sees a hyperlink as evidence of trust within non-spam web, which enables them to detect spam website using web links. Topic-sensitive PageRank is a method to calculate the significance of a web page for specific topic. Sixteen independent topics such as "Health" and "Sports" are prepared in advance using the top categories of Open Directory Project (ODP)[2]. The method calculates scores of a page for each of the topics separately. Our method to calculate impact is also based on biased PageRank. Although thinking about the distance between "Health" and "Sports" does not make much sense, we can introduce the concept of distance between these tempo-spatial topics. The orderly continuity of topics plays an important role in our proposed method.

Wikipedia is a large scale, multilingual, and web based encyclopedia. It has over 3.5 million articles (as of January, 2011) that cover widespread fields. Wikipedia's excellent feature is not just its wide range of fields, but also its organized systematic link structures. It can be used as a very useful corpus for knowledge extraction [4, 3].

There are several environments for the visualization of a historical event sequence. Google Maps [3], a map service on the Web, provides several APIs that can be easily used to create many kinds of geographical Web services. Stoev et al. [18] provides a visual environment especially suited for the historical events. In this paper, the distribution of the geographical impact of a historical entity is visualized with this service.

There are several related techniques and applications that are proposed in the research field of Geographic Information Retrieval (GIR), which is defined by Larson [10, 11]. Strötgen et al. [19] introduced a method for extracting spacio-temporal information from a single document.

In this paper, we divide the surface of the earth in grid shape using the Geohash algorithm, and each grid are used to calculating the spatial impact of historical entity. Martins et al. used the algorithm to extend an XSLT/XQuery engine with additional functions for processing spatial information [2].

Information extraction is another research area related to our work. In particular, Web information extraction has become a large research field. For example, Paşca proposed methods to extract factual information from a large Web corpus [17] and query logs [16]. Although most of the research studies focus on *current* fact extraction, some focus on *historical* fact extraction. Garera et al. [5] proposed a method for biographic fact extraction such as "birthdate", "occupation", "nationality", and "religion". BioSnowball by Liu [12] is a method for gathering and integrating biographical facts. Their method can construct Wikipedia-style pages for any person by aggregating the extracted facts. The values of the attributes of a person can change over time. When extracting the values from the Web, such changes can also be considered [15]. For temporal information extraction, not only the past events but also the future events can be extracted. Jatowt et al. [7] proposed a method for future event extraction from news articles or Web pages. They mainly focused on a method for extracting expressions that predict the future.

3. PROBLEM DEFINITION

It is difficult to compare the impact of a historical person on another one, however, it is surely not easy to compare person with person. For example in an academic domain. We could determine who is a more important scientist by simply counting the number of times all scientist's papers are refereed to by others. But using a paper count as importance determinant does not work when it comes to comparing people with different occupations such as a scientist and a novelist. Therefore, there is a need of some kind of

[2]Open Directory Project: http://www.dmoz.org/

[3]Google Maps: http://maps.google.com/

universal way as a importance determinant to evaluate historical entities.

In the remainder of this section, the definition of historical entity and the concept of impact and significance are described.

3.1 Historical Entity

History is composed of more than just people or events. Many things around us have their historical aspects. For example, although a concept "arithmetic number" looks irrelevant to history, it has its own history of having being discovered. In such perspective, we can regard most of things are relevant to history. In our work, we define the *historical entity* as follows.

Historical entity: An entity must exist, and be distinguishable from other entities. All entities are historical entities, even though they no longer exist.

3.2 Tempo-Spatial Impact of Historical Entity

The impact of the historical entity will vary by facet being considered. For example, if we consider place and time as the facet, *Napoleon* had a major impact in Europe in his era. On the other hand, he had almost no impact in Japan at that time, and he has a little impact in Europe currently.

There are several ways to consider the temporal aspect of historical entity. One way is to extract and count the temporal representation from wiki text and consider it as temporal aspect of corresponding historical entity.

Another way is to use the "Years" Wikipedia categories. This is the portal of the temporal categories, and it has many specific subcategories. For example, "2000" is the child of "Years", and, in addition, categories for 2000 are registered as a subcategory of "2000" recursively.

3.3 Significance of Historical Entity

Significance of historical entity is used to compare them with each other. Each historical entity has its significance as scalar value, although impacts are given as vector value.

3.4 Dataset

Wikipedia offers free copies of all available content to interested users [4]. These snapshots are usually provided monthly. We utilize the data dumped at October 11, 2010.

3.4.1 An Article is a Historical Entity

In this study, we try to represent a historical entity by a Wikipedia article. Then, what kind of articles is appropriate for representing historical entity? As we discussed in Section 3.1, there is a huge variety of historical entities. For instance, the Wikipedia article of "zero" contains the topic about the history of the item itself. Therefore, we decided to consider all Wikipedia articles as historical entities.

In our approach, the significance of historical entity is mainly calculated by using Wikipedia link structure, particularly inbound links, but the temporal distance among these entities is also used. Therefore, all Wikipedia articles contained in the dataset require their own temporal information to be used for measuring the relative distance with the other article. To satisfy the condition, we focus on the "Years" Wikipedia category described in above. We utilize all articles belonging to the category tree as the dataset. In total that were 1,424,616 articles.

[4]http://download.wikipedia.org/enwiki/

3.4.2 Correct Answer Set

We made a correct answer set to evaluate the ranking result. To extract important historical entities impartially, we counted the word that appears in the 11 world history textbooks and extracted words which at least commonly appear in 6 textbooks as the important word in the history. About 3,000 words satisfy the condition. 1,043 articles remained as a result of extracting the Wikipedia article corresponding to these 3,000 words from the data set. Then, these 1,043 articles are used as a correct answer set in this experiment.

4. COMPUTING IMPACT OF HISTORICAL ENTITY

4.1 Link Structure Analysis

In our approach, initial tempo-spacial information of each historical entity is propagated through Wikipedia links in the similar manner as PageRank. PageRank is a method to compute importance of web pages using web link structure. The main idea behind PageRank is that a Web page is important if several other important Web pages link to it. This means that if page u has a link to page v, then the link implicitly confers some importance to v. Then, how should we represent the value of conferred importance? Let $r(u)$ represent the degree of importance of page u, and let F_u represent the set of pages linked by page u. We can simply assume that all links are equal, so the link (u, v) confers $r(u)/|F_u|$ units of importance from page u to page v.

Since $r(u)$ is also recursively determined by r of pages that point to u, usually PageRank algorithm is computed with power method. If N is the number of pages, we assign the initial value $1/N$ to all pages. B_v is the set of pages that points to v. This simple idea leads to the following equation.

$$r_{i+1}(v) = \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} \tag{1}$$

We continue the iterations until all $r(v)$ stabilize within some threshold, although the convergence of the recursion depends on the link structure [14]. Like in a conventional form of PageRank algorithm, we add a damping factor, α, to the rank propagation to guarantee the convergence. We define a new recursion, in which we add the constant value $(1 - \alpha)/N$ to all pages:

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + \frac{1 - \alpha}{N} \tag{2}$$

This modification improves the quality of PageRank by introducing the damping factor α, as well as gurantees the convergence to a certain value for all pages.

The difference between original PageRank and biased PageRank is the added value $(1 - \alpha)/N$ described above. In particular, let C_j be the set of pages in a certain category c_j. We introduce a new damping factor b_{vj} for the page v:

$$b_{vj} = \begin{cases} 0 & v \notin C_j \\ \frac{1}{|C_j|} & v \in C_j \end{cases} \tag{3}$$

Then the biased PageRank is represented as follows:

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{r_i(u)}{|F_u|} + (1 - \alpha) b_{vj} \tag{4}$$

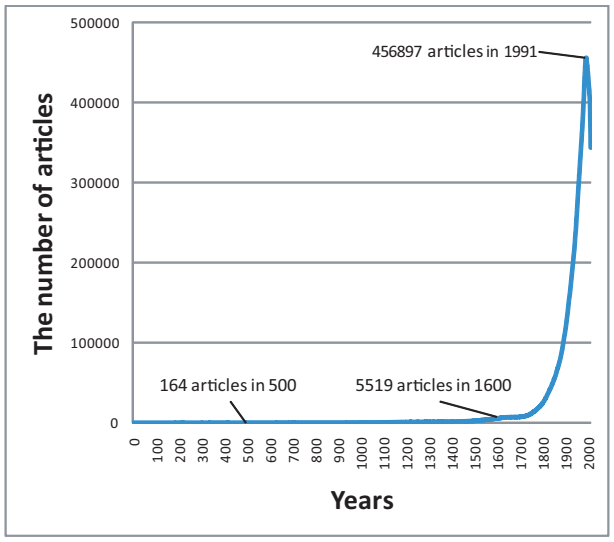

Figure 1: Relation between years and the number of included articles.

In this recursion, pages in the same category and pages that are closely related get high scores.

Note that $|C_j|$, the number of articles, is different for each category c_j.

The relation between the size of $|C_j|$ and temporal category c_j is shown in Figure 1. The number of articles that refer to recent years is increasing rapidly. For this reason, we can not compare the value of the different categories directly. We use a rank within each category to reduce the impact of the size of categories.

4.2 Computing Impact Based on Wikipedia Link Structure

We pay attention to the historical relation between two entities. When dealing with a Wikipedia article as a historical entity, how can we handle the relationship?

In this study, we focus on the relationship of hyperlinks reference between Wikipedia articles. Let an article v be referred to by another article u. Considering the Wikipedia link structure, the anchor text of the link always corresponds to the linked article. All links are created by Wikipedia author group, and according to their editing policy, even if the title of the other article appears in the text, if the context of the sentence is not related to it, the title term will not be marked up as a link. Thus once a link from u to v appears, it indicates strong connection from u to v. Therefore if u points to v, we can regard that v might affected u. On the other hand, this kind of link says nothing about the affection from u to v. For example, if the article of *Einstein* links to the article of *Newton*, *Einstein* shall be deemed to be influenced by *Newton*, but not vice versa.

4.2.1 Temporal Impact

For example, *Newton* was an important figure in 1900, though he had died about two hundred years ago, he may have influenced people and events in that years. This means that importance in a certain period is affected by how the historical entity influences the other entities in the period. We call this kind of importance temporal impact. That is, the impact of the historical entity in a period is determined as follows:

- A historical entity which has major impact to historical entities in a category is important in that period.

The impact of affected entities is also remarkable. Therefore we make the following assumption.

- Historical entity that influenced other important entities in the period has impact in the period.

For instance, *Einstein* is important in the 20th century if it had a big influence on other historical entities belonging to the category of the 20th century. Moreover, if *Einstein* had received a big influence from the *Newton*, *Newton* also has a impact in the 20th century.

The first step in our approach according to the assumption is to calculate a set of biased PageRank scores using a set of temporal topics for all entities. For the biasing factor C_j shown in Equation 3, we use "Years" Wikipedia category. This is the portal of the temporal categories, and it has many specific subcategories. Sometimes an article belongs to several temporal categories. For example, the article of *Newton* belongs to both "1643 births" and "1727 deaths". Then, we assume that each historical entity has two attributes such as "begin year" and "end year", and the oldest year is substituted for "begin year" and latest one is substituted for "end year". For *Newton*, the "begin year" is 1643 and "end year" is 1727.

Note that, these properties can get exactly the same amount for some historical entities. For example, these properties for the article correspond to the 2008 Summer Olympics are 2008.

As topic-sensitive PageRank uses 16 topics $c_1 \cdots c_{16}$ with the ODP, 201 topics $c_0 \cdots c_{200}$ are created with the Wikipedia category in our approach. Each topic c_j represents a decade from $j \times 10$ to $j \times 10 + 9$. For example, c_{200} means period between 2000 and 2009. C_{200} is composed of the articles belonging to Wikipedia categories such as "2000", "2001", ... "2009", and their subcategories.

Since we assume that a Wikipedia link between historical entities represents an impact propagation, biased PageRank score with C_j means the level of impact in the corresponding decade for each historical entity. Figure 2 demonstrates the evolution of the impact of *Leonardo da Vinci*, *Isaac Newton*, and *Albert Einstein* in between C_{100} and C_{200}. The graph indicates that the impact level of *Leonardo da Vinci* started climbing from the 15th century, peaking at the boundary of the 15th century and the 16th century. After that the value decreased rapidly. This does not mean, however, that he had immediately lost the impact after his death. According to Figure 1, the value of $|C_j|$ increases as time goes by. Therefore, the raw values cannot be compared directly even though the value in c_{135} and c_{170} are close to each other. The evolution of the rank of *Leonardo da Vinci*, *Isaac Newton* and *Albert Einstein* is shown in Figure 3. The graph can thus be used to predict that he is still maintaining the impact in these days.

4.2.2 Spatial Impact

We divide the surface of the earth in grid shape with the Geohash algorithm. Originally, the algorithm was invented by Gustavo Niemeyer for the web service named geohash.org

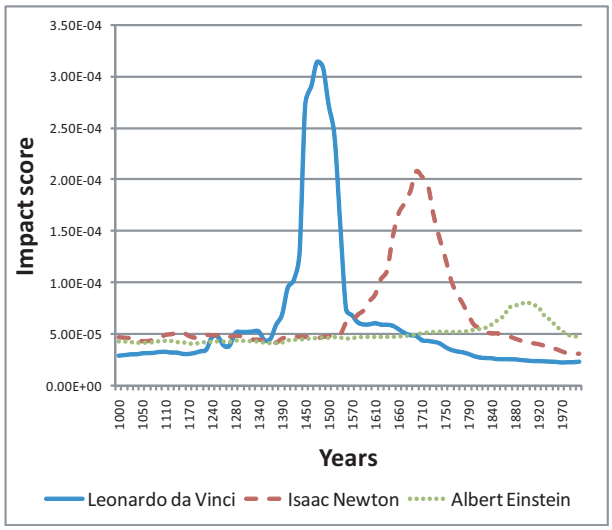

Figure 2: The evolution of the importance of *Leonardo da Vinci*, *Isaac Newton*, and *Albert Einstein* after year 1000.

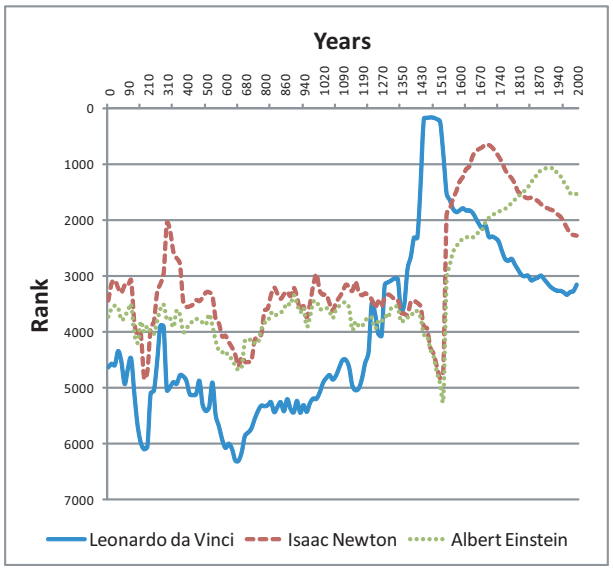

Figure 3: The evolution of the impact rank of *Leonardo da Vinci*, *Isaac Newton*, and *Albert Einstein*.

[5]. Geohash is a latitude/longitude geocode system and a hierarchical spatial data structure which subdivides space into buckets of grid shape.

We divided Europe into grids as shown in Figure 4. In this hierarchy, each grid is represented by three characters. For example, the grid surrounded by a red line, including the Paris, is "u09", blue grid including London is "u10", orange frame including Berlin is "u33", and so on. According to the properties offered by Geohash, both two coordinates that have common prefix are in the same grid represented by the shared part. This means that all coordinates whose Geohash code starts with "u09" exist in the red grid shown in Figure 4. In our study, the spatial categories are represented by these grids. We extract latitude and longitude information from each Wikipedia article. Some Wikipedia articles include this kind of coordinate information. The coordinate of *Arc de Triomphe* is latitude $48°87'N$ and longitude $2°29'E$, and this kind of information is expressed with a *coord tag* like `{coord|48.87|N|2.29|E}` in the wiki text. Then we convert the obtained location information to the Geohash code. For the above example, the coordination of *Arc de Triomphe* is converted to "u09wh1pnt2". Since this starts with "u09", *Arc de Triomphe* is understood to be located in the red grid. In other words, the historical entity of *Arc de Triomphe* is included in C_{u09}. Similarly, the Geohash code of the *Eiffel Tower* is "u09tunqu1x", so the historical entity of the landmark is also included in C_{u09}.

When we calculate the biased PageRank algorithm using the C_{u09}, the given score for a Wikipedia article is the impact of the historical entity around *Paris*.

The results of entities: *Napoleon I*, *Albert Einstein* and *Berlin Wall* are discussed below.

Figure 5 demonstrates the distribution of the geographical impact of *Napoleon I*. The value is widely distributed in Europe, mainly France, where he ascended as the emperor. Moreover, we believe the high value on the route from France to Russia indicates the march of *Napoleon* troops to Russia in 1812. *Napoleon* had also significant impact as the history of Poland as he helped in establishing Duchy of Warsaw.

Figure 6 indicates the results of *Albert Einstein*. It is known that *Einstein* was born in Germany, and worked at the patent bureau in Switzerland when he proposed the Special relativity in 1905. A high value is shown from Germany to Switzerland in the figure.

Figure 7 explains the results of the *Berlin Wall*. The influence not only exists strongly around Germany but also has a high impact on Eastern Europe as compared with Western Europe.

5. EVALUATING SIGNIFICANCE OF HISTORICAL ENTITY

Suppose that *Isaac Newton* had much influence on both *Robert Hooke* and *Albert Einstein*, then which influence should contribute to the significance of *Newton*? One idea is that influence on *Einstein* is more important than that on *Hooke*, because its influence goes beyond different era. *Newton* and *Hooke* were contemporary persons, but *Newton* and *Einstein* lived in different time periods. According to this idea, we propose a hypothesis as follows:

- A historical entity is significant if it influences many other entities that are far from it temporally.

In this section, we propose several techniques to evaluate significance of a historical entity.

Section 5.1 describes the proposed approach using temporal impacts of a historical entity discussed in Section 4.2.1. Several other baseline approaches are described in Section 5.2.

5.1 Evaluate Significance Using Temporal Impact

There are several ways to evaluate the significance of a single historical entity using its temporal impact based on

[5]geohash.org: http://geohash.org/

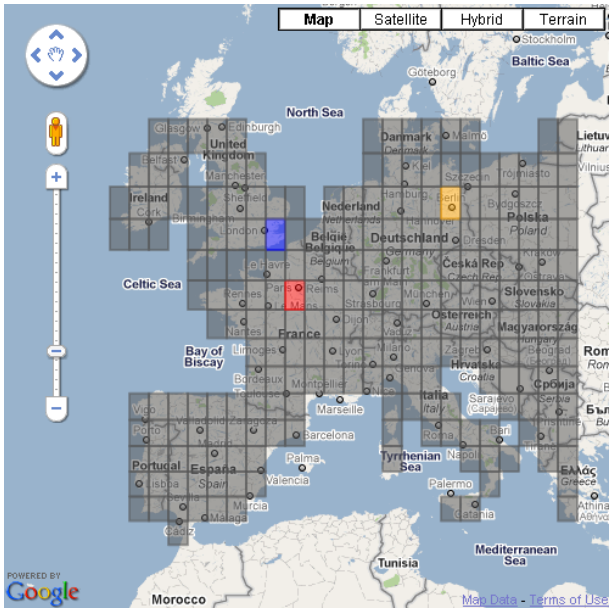

Figure 4: Geo-spatial categories. Each grid shape represents a category.

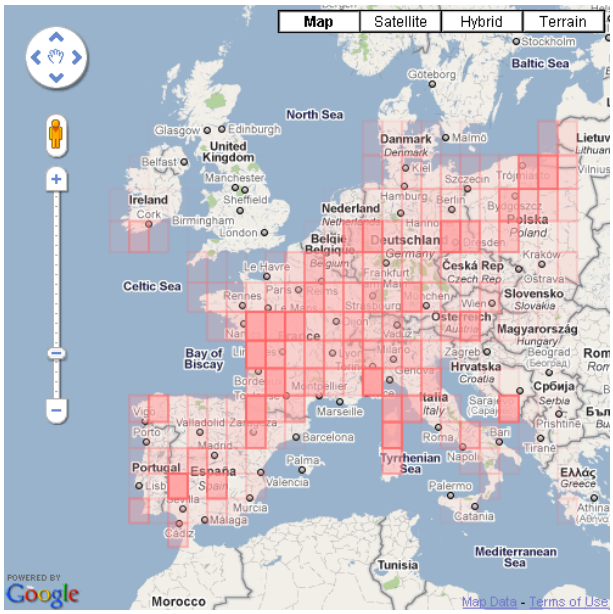

Figure 5: *Napoleon I*

the hypothesis. One way is to regard the sum of temporal impacts as a significance. Though this approach may have a tendency for a historical entity which has big impacts in various era to get a high significance, a one-shot entity, which has a huge impact in only a few era, also can get a high significance.

Another approach considers the temporal distance from the **peak** of an entity. When an impact is further from the peak temporally, this approach gives bigger importance for it. One of the most serious problems of this approach is the indefiniteness of the concept of the peak of a historical entity. Speaking of the historical person, the peak of the person may be the moment when the person was most prominent in his or her life. However how can we decide the moment?

In this paper, we utilize the peak of given temporal impacts for the concept. For *Leonardo da Vinci*, *Isaac Newton* and *Albert Einstein*, the peak is c_{149}, c_{170} and c_{190} respectively (See Figure 2).

The concept of peak is computable for all entities by this approach, and this is an important advantage.

Let e_{peak} the temporal center of a historical entity e and let e_{c_j} represent the impact at c_j. Then $s(e)$, the significance of e, is calculated as follows:

$$s(e) = \sum_{c_j} f(c_j - e_{peak}) e_{c_j} \tag{5}$$

The function f takes a distance between an impact and the peak of entity. In this paper, we define three types of f as follows:

$$f_{linear}(d) = |d| \tag{6}$$

$$f_{log}(d) = log(|d| + 1) \tag{7}$$

$$f_{pow}(d) = d^2 \tag{8}$$

We conduct experiments for each function and show their results in Section 6.

5.2 Baseline Approaches

When dealing with a Wikipedia article as a historical entity, it is easy to consider the length of the article as straightforward a importance measure of a corresponding entity. Generally speaking, articles about the well known subjects tend to have substantial content, so the length of article may be in proportion to the significance of its described object. But, because of the policy of the Wikipedia project, it is possible to arbitrarily lengthen the content, as long as it satisfies neutrality and verifiability of the content description, regardless of the significance of the subject. Therefore, measuring the significance of an article by its length does not have enough robustness. For such reason, in this study, we do not utilize the volume of the contents to evaluate the significance of the historical entity.

In the remainder of this section several other naive approaches to evaluate the significance of a historical entity are described.

5.2.1 *Inbound Link*

In this study, the number of inbound links of a historical entity represents the number of historical entities which the entity influenced. This approach does not care the temporal distance between historical entities.

5.2.2 *Distance Inbound Link*

As we have seen in Section 4.2.1, when each entity has "begin year" and "end year", the temporal distance between two historical entities can be determined. For two historical entities v and u, we define four distance functions $d(u, v)$ as follows (See Figure 8). Let n_{begin} and n_{end} represent "begin year" and "end year" of an entity n respectively, and note that $n_{begin} \leq n_{end}$ for all n.

$$d_{begin}(u, v) = |v_{begin} - u_{begin}| \tag{9}$$

$$d_{end}(u, v) = |v_{end} - u_{end}| \tag{10}$$

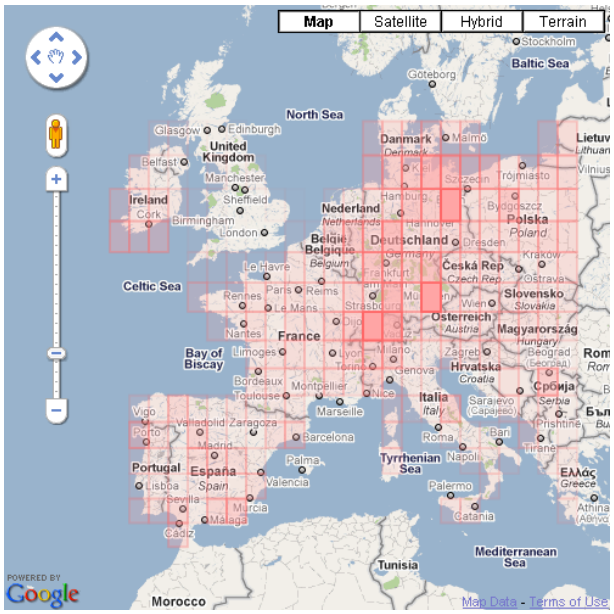

Figure 6: ***Albert Einstein***

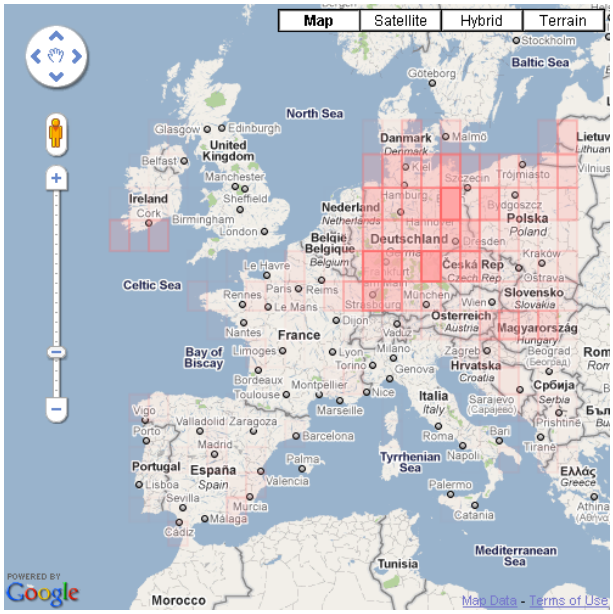

Figure 7: ***Berlin Wall***

$$d_{average}(u,v) = |\frac{v_{end}+v_{begin}}{2} - \frac{u_{end}+u_{begin}}{2}| \quad (11)$$

$$d_{nearest}(u,v) = \begin{cases} 0 & v_{begin} \leq u_{begin} \leq v_{end} \\ 0 & u_{begin} \leq v_{begin} \leq u_{end} \\ v_{end} - u_{begin} & v_{end} \leq u_{begin} \\ u_{end} - v_{begin} & u_{end} \leq v_{begin} \end{cases} \quad (12)$$

We are also able to use the peak of each entity described in Section 5.1 to decide the distance.

$$d_{peak}(u,v) = |u_{peak} - v_{peak}| \quad (13)$$

Note that d_{peak} is computable for all entities unlike four Wikipedia category base approaches. For each distance function d, significance of historical entity v is calculated as follows:

$$s(v) = \sum_{u \in B_v} f(d(u,v)) \quad (14)$$

Using a random sampling, we selected 10,000 links connecting two articles in the database. For these links, the ratio of d_{begin}, d_{end}, $d_{average}$, $d_{nearest}$ and d_{peak} are indicated in Figure 9. It can be seen from the graph that a Wikipedia article has a tendency to refer to other articles which are near to it as a time wise.

On the other hands, Figure 10 demonstrates the ratio of these five distance functions between an entity in the dataset and an entity in the correct answer set. The ratio of $100 < d$ rises obviously compared with Figure 9, and this can be taken to mean that the historical entity included in the correct answer set has a tendency to be linked by various ages compared with the historical entity included in the dataset. We believe that this result supports the our hypothesis that the influence given to the historical entity in a different age is more historically important than to the temporally close one.

5.2.3 *Standard PageRank*

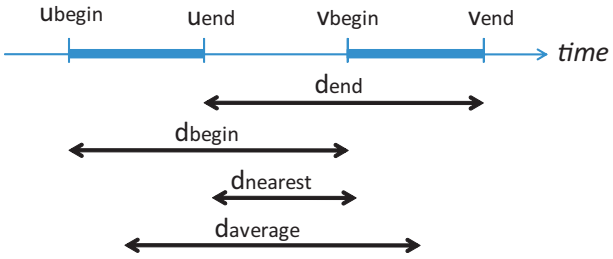

Figure 8: Four kinds of Wikipedia "Years" category base temporal distances between two historical entities u and v.

We focus on how one historical entity influenced other entities, and use that for a significance evaluation. In other words, we think that the historical entity that gave much influence to others is considered to be very significant, and also that even if one entity gave much influence to other entities but these other entities are not significant, then, the influencing entity is considered to have low significance. Therefore we make a following assumption.

- Historical entity that influenced other significant entities is significant.

To calculate such a recursive significance characteristics, we use the PageRank algorithm. That is we compute the PageRank score for each articles using Wikipedia link structure, and then utilize these score as the significance of corresponding historical entity.

5.2.4 *Distance PageRank*

The standard PageRank approach can be improved using the assumption described in the begining of Section 5. We consider the weight of a link as the temporal distance between two entities connected by the link. Let $d(u,v)$ represent the weight of a link from u to v. When the amount of

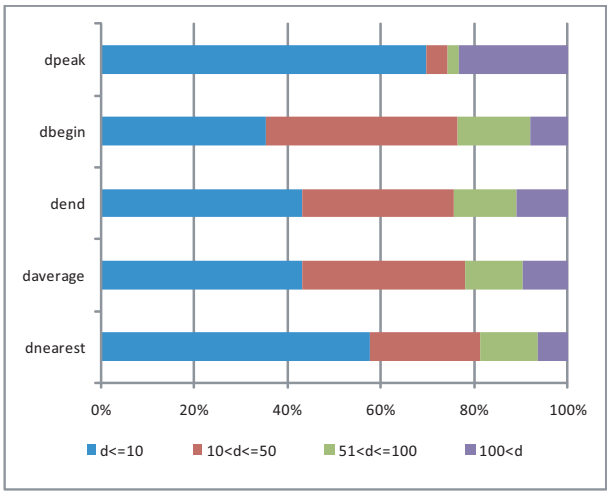

Figure 9: Ratio of result of the four distance function.

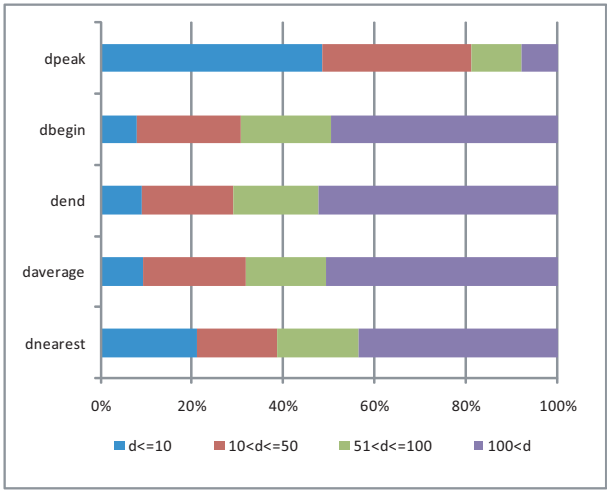

Figure 10: Ratio of value of four distance of links point to item of the correct answer set.

propagated importance is in proportion to weight of itself, the equation 2 is transformed as follows:

$$r_{i+1}(v) = \alpha \sum_{u \in B_v} \frac{d(u,v) r(u)}{\sum_{w \in F_u} d(u,w)} + \frac{1-\alpha}{N} \qquad (15)$$

This impementation is also used in TextRank [13] and VisualRank [8].

6. EXPERIMENT

We conducted several experiments using the dataset and the correct answer set described in Section 3.4 and Section 3.4.2 respectively. These experiments were carried out to determine the effects of twenty-one kinds of significance evaluating technologies described in Section 5. Nine combinations of proposed methods were estimated: there are three kinds of weight function f, equation 6, 7 and 8, and we computed temporal impacts of historical entities with three kinds of damping factor α, 0.15, 0.5 and 0.85 in equation 4. The baseline methods of distance inbound link and distance PageRank were calculated with five kinds of distance function d described in Section 5.2.2.

Table 4: The rank of *Napoleon I*, *Leonardo da Vinci*, *Lobert Hooke*, *Isaac Newton* and *Albert Einstein* using proposed method with $\alpha = 0.85$.

	f_{linear}	f_{log}	f_{pow}
Napoleon I	11	24	8
Leonardo da Vinci	433	511	421
Lobert Hooke	2137	2640	2195
Isaac Newton	99	150	91
Albert Einstein	66	192	34

Table 1 demonstrates the precision and recall of sorting results. For distance inbound link and distance PageRank, the best results are displayed. Speaking of the proposed methods, the larger the damping factor α is, the better the result is, and the selection of weight function f is not so important. The result of the proposed methods was better than the result of the baseline methods. The distance inbound link and distance PageRank are better than inbound link and standard PageRank respectively. We believe that these results support the hypothesis that historical entity is significant if it influences many other entities that are far from it temporally.

Various kinds of historical entities are included in the dataset. We used Wikipedia categories related to human to squeeze the dataset and the correct answer set to the historical person. The squeezed dataset and the correct answer set were composed by 474,680 entities and 392 entities respectively. The result of precision and recall for these new dataset and the correct answer set is shown in Table 2. Comparing with the previous case, the results of baseline methods are very improved though the results of proposed methods are almost changeless. This means that non-human historical entity could not be extracted effectively by these baseline methods. Table 3 demonstrates the top 10 results. The historical person who is written by bold font is included in the correct answer set. It is understood that the person who belongs to a specific group (e.g. President of the States, Pope and Roman Emperor, etc.) takes a high order from this result. Because a historical entity that represents the group itself is referred from much variety of periods, a historical person referred by the entity may have a tendency to get a high significance. For example, there is *President of the United States* in the Wikipedia and the entity is referred by many other entities in the history of the United States. Finally, the rank of four historical people mentioned in this paper is shown in table 4. According to the proposed methods, *Napoleon* is the most significant among five people, and "*Einstein* is a little greater than *Newton*" is our answer for the question "Which is greater, *Albert Einstein* or *Isaac Newton*?".

7. CONCLUSION AND FUTURE WORK

We proposed several approaches to compute the some kinds of aspects of a historical entity. Our method attempts to sort the history of humankind according to its historical importance.

Approaches to compute the impact of a historical entity are proposed. Temporal distance among the historical en-

Table 1: Precision and Recall of sorting all historical entities.

Name	@10	@100	@1000	@10000	@100000
proposed f_{linear} 0.15	0.3 | 0.00301205	0.31 | 0.0311245	0.113 | 0.113454	0.0274 | 0.2751	0.0038 | 0.381526
proposed f_{log} 0.15	0.4 | 0.00401606	0.28 | 0.0281124	0.092 | 0.0923695	0.028 | 0.281124	0.00379 | 0.380522
proposed f_{pow} 0.15	0.2 | 0.00200803	0.27 | 0.0271084	0.108 | 0.108434	0.0269 | 0.27008	0.00378 | 0.379518
proposed f_{linear} 0.5	0.3 | 0.00301205	0.31 | 0.0311245	0.138 | 0.138554	0.0454 | 0.455823	0.00687 | 0.689759
proposed f_{log} 0.5	0.4 | 0.00401606	0.28 | 0.0281124	0.136 | 0.136546	0.0449 | 0.450803	0.00713 | 0.715863
proposed f_{pow} 0.5	0.2 | 0.00200803	0.27 | 0.0271084	0.128 | 0.128514	0.0424 | 0.425703	0.00633 | 0.635542
proposed f_{linear} 0.85	0.4 | 0.004016	0.29 | 0.029116	0.182 | 0.182731	0.0681 | 0.683735	0.00956 | **0.959839**
proposed f_{log} 0.85	0.5 | 0.005020	0.33 | 0.033132	0.17 | 0.170683	0.0663 | 0.665663	0.0096 | **0.963855**
proposed f_{pow} 0.85	0.4 | 0.004016	0.25 | 0.025100	0.19 | 0.190763	0.065 | 0.65261	0.00948 | **0.951807**
inbound link	0.4 | 0.004016	0.28 | 0.028112	0.077 | 0.077309	0.0204 | 0.204819	0.00329 | 0.330321
$d_{average}$ link	0.5 | 0.005020	0.19 | 0.019076	0.11 | 0.110442	0.0274 | 0.2751	0.0038 | 0.381526
PageRank	0 | 0	0.09 | 0.009036	0.08 | 0.080321	0.0205 | 0.205823	0.00282 | 0.283133
d_{peak} PageRank	0.1 | 0.002551	0.05 | 0.012755	0.078 | 0.19898	0.019 | 0.484694	0.00269 | 0.686224

Table 2: Precision and Recall of sorting all historical people.

Name	@10	@100	@1000	@10000	@100000
proposed f_{linear} 0.85	0.6 | 0.0153061	0.31 | 0.0790816	0.147 | 0.375	0.0343 | 0.875	0.00387 | 0.987245
proposed f_{log} 0.85	0.6 | 0.0153061	0.26 | 0.0663265	0.133 | 0.339286	0.0352 | 0.897959	0.0039 | 0.994898
proposed f_{pow} 0.85	0.7 | 0.0178571	0.37 | 0.0943878	0.148 | 0.377551	0.0324 | 0.826531	0.00385 | 0.982143
inboud link	0.4 | 0.0102041	0.28 | 0.0714286	0.077 | 0.196429	0.0204 | 0.520408	0.00329 | 0.839286
$d_{average}$ link	0.5 | 0.0127551	0.19 | 0.0484694	0.11 | 0.280612	0.0274 | 0.69898	0.0038 | **0.969388**
PageRank	0 | 0	0.09 | 0.0229592	0.08 | 0.204082	0.0205 | 0.522959	0.00282 | 0.719388
d_{peak} PageRank	0.1 | 0.00255102	0.05 | 0.0127551	0.078 | 0.19898	0.0189 | 0.482143	0.00269 | 0.686224

tities plays an important role in our methods, so we utilize the "Years" category on Wikipedia to determine the temporal information for each historical entity to calculate temporal impacts. There are, however, mainly two problems with the approach. First one is the issue of the coverage of Wikipedia articles. The number of articles belonging to the category is 1.4 million, though Wikipedia offers currently more than 3.5 million articles. This means that we can not partially take advantage of the quantity of Wikipeida. The other is the issue of the inaccuracy of prediction. For example, *Japan* belongs to "States and territories established in 660 BC" category, so the entity's "begin year" and "end year" is set to 660 BC. To address these issues, we are going to use natural language processing. In addition, our method makes it possible to compare the influence of different people on the same category, and the influence of same person on the different categories, but still impossible to compare the influence of different persons on the different categories. For example, we can not compare "*Newton* in 1700" and "*Einstein* in 1900". We also are working on this issue.

A hypothesis for the significance of historical entity and the way to evaluate the significance using the hypothesis are also proposed. Our approach focuses on the diversity and the contiguousness of the temporal impacts of each historical entity. We estimated several approaches to evaluate the significance of historical entity by conducting some experiments, and the result of these experiments supported the hypothesis. As we discussed in previous section, there is a tendency that a historical entity which belongs to particular category (e.g. country and the President of the United States, etc.) gets a especially good position. We believe that the diversity of the result is useful to summarize the mankind history. To address the issue, we will have to find a historical entity which plays a *hub* role such as *President of the United States*. The advantage of HITS algorithm [9] could be useful.

8. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of the 7th International Conference on World Wide Web (WWW 1998)*, pages 107–117, April 1998.

[2] N. F. Bruno Martins and J. Borbinha. Complex data transformations in digital libraries with spatio-tmporal information. In *Proc. of the 11th International Conference on Asia-Pacific Digital Libraries (ICADL 2008)*, pages 174–183, December 2008.

[3] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 708–716, June 2007.

[4] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1606–1611, January 2007.

[5] N. Garera and D. Yarowsky. Structural, transitive and latent models for biographic fact extraction. In *Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, pages 115–124, June 2009.

[6] T. H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data*

Table 3: Sorting result of historical people. Bold person is included in correct answer set.

	proposed f_{linear} 0.85	proposed f_{log} 0.85	proposed f_{pow} 0.85	inboud link	$d_{average}$ inboud link
1	**Augustus**	Charlemagne	Augustine of Hippo	**George W. Bush**	**Augustus**
2	Pope John Paul II	**Augustus**	Pope John Paul II	Bill Clinton	**William Shakespeare**
3	Charlemagne	**Muhammad**	**Plutarch**	Barack Obama	**Saint Peter**
4	Ptolemy	Pope John Paul II	Ptolemy	**Ronald Reagan**	John Chrysostom
5	**Plutarch**	**Augustine of Hippo**	**George W. Bush**	Bob Dylan	Jerome
6	**Saint Peter**	Diocletian	**Saint Peter**	Robert Christgau	**Muhammad**
7	**George W. Bush**	**Plutarch**	**Adolf Hitler**	**Adolf Hitler**	Athanasius of Alexandria
8	**Augustine of Hippo**	**Saint Peter**	**Napoleon I**	**John F. Kennedy**	Gregory of Nazianzus
9	**Muhammad**	**Justinian I**	Charlemagne	Michel Jackson	Hilary of Poitiers
10	Diocletian	Ptolemy	**Augustine of Hippo**	Elvis Presley	**Pope Gregory I**

Engineering (TKDE 2003), 15(4):784–796, July/Aug 2003.

[7] A. Jatowt, K. Kanazawa, S. Oyama, and K. Tanaka. Supporting analysis of future-related information in news archives and the web. In *Proc. of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2009)*, pages 115–124, June 2009.

[8] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30:1877–1890, 2008.

[9] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] R. R. Larson. Geographic information retrieval and spatial browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, pages 81–124, April 1996.

[11] R. R. Larson and P. Frontiera. Geographic information retrieval (gir): Searching where and what. In *Proc. of the the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004)*, page 600, July 2004.

[12] X. Liu, Z. Nie, N. Yu, and J.-R. Wen. Biosnowball: Automated population of wikis. In *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 969–978, July 2010.

[13] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proc. of the 2004 Conference of the Empirical Methods in Natural Language Processing*, pages 404–411, July 2004.

[14] R. Motwani and P. Raghavan. Randomized algorithms. *ACM Computing Surveys (CSUR)*, 28(1):33–37, March 1996.

[15] S. Oyama, K. Shirasuna, and K. Tanaka. Identification of time-varying objects on the web. In *Proc. of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*, pages 285–294, June 2008.

[16] M. Paşca. Organizing and searching the world wide web of facts – step two: Harnessing the wisdom of the crowds. In *Proc. of the 16th International Conference on World Wide Web (WWW 2007)*, pages 101–110, May 2007.

[17] M. Paşca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. Organizing and searching the world wide web of facts – step one: the one-million fact extraction challenge. In *Proc. of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, pages 1400–1405, July 2006.

[18] S. L. Stoev, M. Feurer, and M. Ruckaberle. Exploring the past: a toolset for visualization of historical events in virtual environments. In *Proc. of the the ACM Symposium on Virtual Reality Software and Technology (VRST 2001)*, pages 63–70, November 2001.

[19] J. Strötgen, M. Gertz, and P. Popov. Extraction and exploration of spatio-temporal information in documents. In *Proc. of the 6th ACM Workshop on Geographic Information Retrieval (GIR 2010)*, pages 1–8, February 2010.

[20] H. G. M. Zoltan Gyöngyi and J. Pedersen. Combating web spam with trustrank. In *Proc. of the 30th International Conference on Very Large Data Bases (VLDB 2004)*, pages 576–587, August 2004.