Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Y. Huang, M. Gertz, J. C. Krumm, J. Sankaranarayanan, and M. Schneider, eds., pages 153–162, Dallas, TX, November 2014.

Uncovering the Spatial Relatedness in Wikipedia

Gianluca Quercini
Supélec E3S
3, rue Joliot Curie, 91190
Gif-sur-Yvette, France
gianluca.quercini@supelec.fr

Hanan Samet
University of Maryland
A.V. Williams Building
College Park, MD 20742, USA
hjs@cs.umd.edu

ABSTRACT

In a previous work we showed that the knowledge of the *spatial reader scope* of a news source, that is the geographical location for which its content has been primarily produced, plays an important role in disambiguating *toponyms* in news articles. The determination of the spatial reader scope of a news source is based on the notion of a *local lexicon*, which for a location l is defined as a set of *concepts*, such as names of people, landmarks and historical events, that are *spatially related* to l. The automatic determination of a local lexicon for a wide range of locations is key to implementing an efficient geotagged news retrieval system, such as *NewsStand* and its variants *TwitterStand* and *PhotoStand*. The major research challenge here is the measurement of the spatial relatedness of a concept to a location. Our previous work resorted to a similarity measure that used the geographic coordinates attached to the Wikipedia articles to find concepts that are spatially related to a certain location. Clearly, this results in local lexicons that mostly include spatial concepts, although non-spatial concepts, such as people or food specialties, are key elements of the identity of a location. In this paper, we explore a set of *graph-based* similarity measures to determine a local lexicon of a location from Wikipedia without using any spatial clues, based on the observation that the spatial relatedness of a concept to a location is hidden in the Wikipedia link structure. Our evaluation on the local lexicons of 1,200 locations indicates that our observation is well-founded. Additionally, we provide experiments on standard datasets that show that SYNRANK, one of the measures that we propose for computing the spatial relatedness of a concept to a location, rivals existing similarity measures in determining the *semantic relatedness* between wikipedia articles.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Search - Information Filtering

General Terms

Algorithms

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Forth Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666398.

Keywords

spatial relatedness, toponym disambiguation, geotagging

1. INTRODUCTION

Suppose that you see the term "Washington" in a news article. There are many possible interpretations including the names of people, monuments, cities, counties, states, etc. The correct interpretation depends on context. In particular, the presence of related terms in the text, such as "Lincoln Memorial", "Vincent C. Gray" and "Wmata", corresponding to the name of a monument, the mayor, and a public transportation system respectively, are useful in narrowing the correct interpretation to "Washington, D.C." The automatic selection, or *disambiguation*, of the correct interpretation of toponyms (i.e., textual specifications of geographical locations) [12, 13] in news articles is central in geotagged news retrieval systems, such as *NewsStand* [14, 24, 26, 27, 30] and its variants *TwitterStand* [8, 11, 29], *STEWARD* [16], and *PhotoStand* [5, 23, 28] which are text-based extensions of the SAND Browser [3, 25].

In a previous work we showed that the disambiguation of toponyms can be effectively achieved by determining the spatial reader scope of a news source [21]. The *spatial reader scope* of a news source, such as a newspaper or a news RSS feed, is the geographical location for which the content of the source is primarily created or, in other words, an audience that is familiar with and interested in news topics about the location and surrounding areas [21]. This means that references to these topics in articles are assumed to be local unless some additional spatial qualification to the contrary is provided. For example, any mention of "Paris" in the articles produced by the Texan newspaper "The Paris News" should be interpreted as a reference to "Paris, Texas", unless other explicit textual evidences point to different interpretations.

The determination of the spatial reader scope of a news source is based on the notion of a *local lexicon*, which for a *location* l is defined as a set of *concepts* which are *spatially related* to l. A concept c, which may denote different types of entities, such as, but not limited to, people, landmarks, food specialties, historical events, movies, songs and companies, is *spatially related* to a location l if c can be unambiguously associated to l. For example, "cheesesteak", "Michael Nutter", "Great Central Fair" and "Philadelphia Museum of Art" are all concepts spatially related to "Philadelphia, PA", because they refer respectively to a food specialty, the mayor, an historical event, and a landmark which unambiguously identify the city of Philadelphia.

The major research challenge for the identification of a local lexicon is the determination of the spatial relatedness of a concept to a location. In our previous work, we resorted to a similarity measure that used the geographic coordinates attached to the Wikipedia

articles to find concepts that are spatially related to a certain location. Clearly, this results in local lexicons that mostly include spatial concepts, although non-spatial concepts, such as people or food specialties, are key elements of the identity of a location.

In this paper, we expand on our previous work and explore a set of graph-based similarity measures to determine the spatial relatedness of a concept to a location, based only on the link structure of Wikipedia, which is generally abstracted as a graph with a node for every article and an edge between any two nodes that are linked in Wikipedia. Importantly, the measures do not use any of the spatial clues provided by Wikipedia. Our rationale is that the spatial relatedness of a concept to a location is already hidden in the Wikipedia link structure. Indeed, articles that describe concepts (e.g., "Eiffel Tower" and "Paris Saint-Germain F.C.") that are spatially related to a location (e.g., "Paris") usually have links to and from the article describing the location. Moreover, for any two articles that describe two spatially related concepts (e.g., "Paris" and "Eiffel Tower"), there are usually many articles that link to both (e.g., "Gustave Eiffel", "Arc de Triomphe", "The Louvre", "Bastille Day").

The following are the key contributions of our paper:

- We propose new graph-based similarity measures that address the limitations of existing measures for the computation of the spatial relatedness of a concept to a location.
- We provide a comparative evaluation of the measures for the creation of a local lexicon of 1,200 locations. Our evaluation indicates that our observation that the spatial relatedness of concepts to locations is hidden in the Wikipedia link structure is well-founded.
- Experiments on standard datasets show that SYNRANK, one of the measures that we propose for the computation of spatial relatedness, rivals existing similarity measures in determining a more general *semantic relatedness* between words.

The rest of the paper is organized as follows. Section 2 reviews existing similarity measures. Section 3 presents basic notation and definitions. Section 4 describes a set of similarity measures that are thoroughly evaluated in Section 5. Section 6 concludes the presentation.

2. RELATED WORK

In this paper our focus is on the extraction of local lexicons of articles from Wikipedia, which requires the ability to measure the relatedness of articles, a problem that has already been addressed by numerous researchers [6, 17, 18, 20, 21, 31, 32]. In the case of our problem, the main question we need to answer is how we measure the spatial relatedness of articles.

Existing relatedness measures can be classified roughly into two groups: text-based and graph-based, which determine the relatedness of two articles by respectively using textual features and the link structure of the Wikipedia graph. Among the text-based measures, the most cited and effective appears to be *Explicit Semantic Analysis* (ESA) [6]. ESA represents any text as a weighted vector (the *interpretation vector*) of Wikipedia articles, each weight representing the relevance of a Wikipedia article to the given text. Based on the assumption that two semantically related texts are represented by similar interpretation vectors, their relatedness score is computed as the cosine similarity of the corresponding interpretation vectors. Although ESA proved to be effective in determining the semantic relatedness of two texts, it generally needs more computational time than graph-based measures. Moreover, the computed relatedness scores vary greatly depending on how the Wikipedia articles are preprocessed to create the interpretation vectors.

Among graph-based measures, the measure proposed by Milne and Witten [17], named WLM, is based on the *normalized Google distance* [1]. The rationale of WLM is that two articles are related if there are many articles that link to both. This measure has two major issues, especially when it comes to evaluating spatial relatedness. First, it excessively penalizes articles with a low indegree [1]. For instance, the spatial relatedness of "La Ruche" (with indegree 17) to "Paris" is measured by WLM as much weaker than the spatial relatedness of "Marseille" (with indegree 3,269) to "Paris", although "La Ruche" is a small neighborhood of "Paris" while "Marseille" is a city located almost 500 miles away from Paris. Second, often two articles with large indegree are likely to have many articles that link to both; for instance, there are 2,491 articles that link to both "Rome" and "Paris" and 2,549 articles that link to both "Berlin" and "Paris", most probably because they are all important European capitals. This leads WLM to determine that "Rome" and "Berlin" are spatially related to "Paris" as much as "Sorbonne" and "Musée du Louvre", which are important landmarks in Paris. Our relatedness measure SYNRANK overcomes both issues, as shown also by our experiments in Section 5. The measure proposed by Ollivier and Senellart [18], referred to as GREEN, is based on Green functions, which are widely used in Markov Chain theory, although they were not intended for NLP-related tasks. GREEN also suffers from the same issues as WLM, but the latter is better at capturing the spatial relatedness of a concept to a location, as our evaluation reveals (Section 5).

Finally, to the best of our knowledge, our previous work is the only one that focuses on the spatial relatedness of a concept to a location [21]. The measure that we proposed makes use of the spatial coordinates that are provided by Wikipedia for the articles describing spatial concepts. Although the method works fine, the resulting local lexicons mostly include spatial concepts; the measures that we explore in this paper address this point.

3. PRELIMINARIES

Wikipedia is the largest encyclopedia available online and, according to a now famous survey [7], rivals *Britannica* in accuracy, even though it can be edited by everyone. Not only is it a source of information for people, but it is also regarded as a relatively structured knowledge base that is commonly used in NLP applications.

In Wikipedia, each *article* has a unique title and is dedicated to a specific *concept*. It usually includes an *introduction*, which highlights some important information, and one or more sections that detail different aspects of the concept. Optionally, an *infobox* is provided containing an overview of the article in a table with a predefined format. Hyperlinks are used between articles to let people quickly browse and discover knowledge related to what they read, while *redirect* and *disambiguation* pages help readers to find articles when they do not know their exact titles.

Each article is assigned to one or more *categories* which are listed at the end of the article and consist of a set of keywords that describe concisely what the article is about. For instance, the article "Barack Obama" is included in categories such as "African-American lawyers", "Presidents of the United States" and "Illinois State Senators". The motivation for categories is to help readers browse the articles, and to this extent they are organized in a hierarchy, as each may branch into *subcategories*, as well as also possibly being included in one or more categories. Both the set of articles and categories can be described by means of a graph, the

[1] *Indegree* of an article: number of links pointing to that article.

Table 1: Statistics in Wikipedia as of October 2010

	Concepts	Spatial concepts	Implicit spatial concepts	Redirect concepts	Disambig. concepts	Links	Categories
English	3,265,081	568,005	629,574	4,497,407	204,737	67,022,955	631,158
German	1,022,944	98,143	275,749	797,149	126,595	24,576,431	94,610
French	955,010	127,750	183,132	1,053,201	63,736	20,758,410	158,186
Italian	700,884	124,034	94,543	367,408	46,729	15,052,156	112,040
Spanish	669,012	104,967	91,160	1,159,790	28,341	14,004,456	126,319

Wikipedia graph and the *category network*.

The Wikipedia graph W is a directed graph with a node v for each article $a(v)$ and a link connecting node v to node w if $a(v)$ has a link to $a(w)$. Also, each link (v, w) in W has two flags to signal whether the corresponding link occurs in the infobox or in the introduction of $a(v)$ respectively. Links in the infobox or the introduction of $a(v)$ are in a prominent position and likely to be more important than the links that occur in the body of $a(v)$. Since every article is devoted to a specific concept, a node in W is called *concept*, has a unique label, given by the title of the corresponding article, and metadata that express important characteristics of the corresponding article, such as its length (in number of characters), whether it is a redirect or a disambiguation page, and the spatial coordinates. The latter are associated with *spatial concepts*, that represent entities (cities, landmarks, events) that have a precise position in the space, as well as with *implicitly spatial concepts*, which denote a particular geographic area though they do not have a fixed position. Examples of spatial concepts are "Paris", "Empire State Building" and "Assassination of Abraham Lincoln", which took place at Ford's Theater in Washington, D.C., while implicitly spatial concepts are "Nicolas Sarkozy" (which implies "France"), "Crab cake" (which denotes the state of Maryland but it can be found also in other states) and "Italian Socialist Party" (which implies Italy). The number of spatial concepts and implicitly spatial concepts in different versions of Wikipedia are tabulated in Table 1.

In the category network C, each node v corresponds to a category $\gamma(v)$ and each link joins a node v with a node w if $\gamma(w)$ is a subcategory of $\gamma(v)$. In this case, $\gamma(w)$ is said to be the *child* of $\gamma(v)$, and $\gamma(v)$ is said to be the *parent* of $\gamma(w)$. If there is a path from category $\gamma(v)$ to category $\gamma(w)$, then $\gamma(w)$ is called a *descendant* of $\gamma(v)$, and $\gamma(v)$ is called an *ancestor* of $\gamma(w)$. Each node v is labeled with the name of its corresponding category $\gamma(v)$, which must be unique, and has a list of concepts that are assigned to $\gamma(v)$. Every element of this list is a reference to a node in W.

4. LOCAL LEXICON CREATION

In this section we describe two methods that given a concept r, referred to as the *root concept*, create the local lexicon $L(r)$ of r. The root concept r denotes a geographic location (e.g., a city). One method uses the *Wikipedia graph* W and is based on a similarity measure that assigns a *score* to the concepts in W, where ideally concepts with a high score are strongly spatially related to concept r and concepts with a low score are weakly spatially related or spatially unrelated to r. The other method uses the Wikipedia *category network*. In the remainder of this section, we give greater details on both methods.

4.1 Using the Wikipedia Graph

A possibility for creating a local lexicon $L(r)$ of r is to compute the spatial similarity score between r and all concepts in Wikipedia, and include in $L(r)$ only those concepts that have a similarity score above a certain threshold. However, this solution can be computationally expensive, because, as detailed in Table 1, the most popular language editions of Wikipedia have a large number of concepts, which increases at an incredibly fast pace over time. As of October 2010, the English Wikipedia had almost 8 million concepts (if we count the redirect and disambiguation concepts). As of May 2014, the number of concepts is almost 11 million, which amounts to an increase of 3,000,000 nodes in 4 years.

To limit the number of concepts for which a similarity score is computed, we observe that most concepts that have a link to or from r (i.e., the *neighbors* of r) are likely to be spatially related to r. Moreover, a concept that has no link to or from r can still be spatially related to r if it is connected to one neighbor of r that is strongly spatially related to r. We consider that those concepts that have no link to r nor to any of its neighbors are unlikely to be spatially related to r at all. Therefore, given a root concept r, we compute the spatial similarity score $S(c, r)$ of a concept c only if:

- c is a neighbor of r, or
- c has a link to or from a concept d which has a mutual link to r.

We now describe the similarity measures that we used to compute $S(c, r)$.

4.1.1 Jaccard-Based Measures

The *Jaccard index* is a well-known similarity coefficient, introduced by botanist Paul Jaccard in 1901 to measure the similarity between two sets A and B [10], and is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad 0 \leq J(A, B) \leq 1$$

That is, the more items shared by sets A and B, the more similar they are. This principle can be applied to the Wikipedia graph W as well, where intuitively two nodes having many common neighbors (i.e., nodes to which they link or they are linked by) are likely to be related. Specifically, the first measure that we propose, called JACC, simply computes the relatedness of two nodes r and c in W by $J(N(r), N(c))$, where $N(i)$ is the set of neighbors of node i in W.

Our evaluation (Section 5) shows that the application of the Jaccard index to the neighbor sets of two nodes is good for measuring the semantic relatedness of two concepts but not their spatial relatedness. In particular, we observed that two concepts tend to share many neighbors when they are related and they are of the same type (e.g., city, landmark, person). On the other hand, two related concepts that do not describe entities of the same type (e.g., a landmark and a city) do not share many neighbors. This explains why, for example, the Jaccard index of the neighbor sets of "Washington, D.C." and "New York City" is higher than the Jaccard index of the neighbor sets of "Washington, D.C." and "White House", even though the second pair of concepts are, from an intuitive point of view, more spatially related than the first pair. We conclude that not all the neighbors of a concept are equally important and that

to measure a spatial relatedness score between a concept c and the root concept r we need to apply the Jaccard index to selected subsets of their neighbors.

We therefore propose another measure called JACCOPT, which is based on the notion of a *kernel* of a concept [21]. The kernel of a concept c, denoted as $K(c)$, is the set of concepts that best describe c and therefore are highly related to it. JACCOPT creates $K(c)$ as a subset of the out-neighbors [2] of c such that a concept $c_1 \in K(c)$ when both the following conditions are satisfied:

1. There is a link directed from c_1 to c.
2. The link from c_1 to c occurs either in the infobox or introduction of c_1's Wikipedia article.

The first condition states that two concepts that mutually link to each other are likely to be spatially related, while the second one acknowledges that this is not always true (e.g., "Paris" and "New York City") and stipulates that one of the two links appear in a prominent position within the Wikipedia article where it occurs (e.g., "Paris" and "Eiffel Tower").

We now measure the relatedness between the root concept r and concept c by looking at the extent to which they link to the same concepts and that these concepts link back to r. We do this by computing the weighted Jaccard index of $K(r)$ and $K(c)$ which is given by

$$JaccOpt(c, r) = w(c, r) \cdot J(K(r), K(c))$$

where the value of $w(c, r)$ is the result of manual tuning designed to give a boost to concepts that are in the kernel of r and concepts that have links to r that occur either in the introduction or in the infobox of their corresponding Wikipedia articles. Specifically, the values of $w(c, r)$ are as follows:

- 3 if $c \in K(r)$ or c has a link to a concept in $K(r)$ or has a link to r that occurs either in the introduction or in the infobox of c's Wikipedia article.
- 2 if c has a link to r in the introduction or in the infobox of c's Wikipedia article and r has no link to c.
- 1.5 if r and c mutually link to each other but $c \notin K(r)$.
- 1 in the remaining cases.

Note that there must not necessarily be a link from r to c nor from c to r for $J(K(r), K(c))$ to be nonzero. As our evaluation shows, JACCOPT is better than JACC when it comes to determining the spatial relatedness of two concepts.

4.1.2 The SYNRANK Relatedness Measure

In this section, we describe a novel relatedness measure, which we call SYNRANK, which evaluates the relatedness of two concepts x and y by capturing characteristics of the inlinks of x and y.

This inlink measure consists of three essential components, $\mathcal{C}_1$–$\mathcal{C}_3$, which capture qualities of related concepts x and y, and are multiplied to form SYNRANK. These components include:

1. $\mathcal{C}_1$: *Pointwise mutual information (PMI)*, which measures how frequently x and y are linked to by Wikipedia concepts, rather than being linked separately;
2. $\mathcal{C}_2$: *Shared inlink boosting*, where concepts that share many common inlinks are deemed more related; and
3. $\mathcal{C}_3$: *Graph distance*, where x and y are deemed more related if the distance between them in the Wikipedia graph is smaller.

Below, we describe these components in greater detail.

Pointwise Mutual Information. Our first component $\mathcal{C}_1$ is based on the notion of *pointwise mutual information* (PMI) [2], which is a measure of association used in information theory and statistics. If X and Y denote two random variables, the pointwise mutual information between two possible outcomes $X = x$ and $Y = y$ is

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

where P is a probability function and $P(x, y)$ is the joint probability distribution function of X and Y. The intuition is that $PMI(x, y)$ has a positive value when x and y are positively correlated and a negative value when x and y are negatively correlated.

We apply these ideas to Wikipedia by considering that X (Y) is a random variable that represents a link from a Wikipedia concept to concept x (y). Here, $P(x)$ ($P(y)$) is the probability that a concept has a link to concept x (y), while $P(x, y)$ is the probability that a concept has links to both x and y. Letting $f(x)$ ($f(y)$) be the number of concepts that link to x (y), $f(x, y)$ be the number of concepts that link to both x and y, and N be the total number of concepts in Wikipedia, then we have that

$$P(x) = \frac{f(x)}{N}, P(y) = \frac{f(y)}{N}, P(x, y) = \frac{f(x, y)}{N}.$$

Substituting into the $PMI(x, y)$ definition above, we have a first relatedness measure:

$$PMI_1(x, y) = \log \frac{N \cdot f(x, y)}{f(x)f(y)}.$$

This formula has a problem. Consider three concepts x_1, x_2 and y, such that

$$\frac{f(x_1, y)}{f(x_1)} \gg \frac{f(x_2, y)}{f(x_2)}$$

Since the concepts that link to both x_1 and y account for a large part of the indegree of x_1, much more than in the case of the concepts that link to both x_2 and y, the pair of concepts (x_1, y) should have a much higher relatedness score than (x_2, y). However, the values of $PMI_1(x_1, y)$ and $PMI(x_2, y)$ tend to be comparable, which is due the presence of the logarithm. We fix this problem by removing the logarithm, which leads to the definition of our first component.

$$\mathcal{C}_1(x, y) = \frac{N \cdot f(x, y)}{f(x)f(y)}$$

Shared Inlink Boosting. Nevertheless, the above solution is not satisfactory, because, as noted by [19], PMI (and therefore $\mathcal{C}_1$) tends to assign high scores to low frequency events, which here means that concepts with low indegree are assigned high scores. For example, if concept x is "Paris" and concept y is "Eiffel Tower", a major symbol of Paris, $\mathcal{C}_(x, y)$ has a lower value than when y is "Tour Carpe Diem", which is a minor landmark in Paris. That is, "Tour Carpe Diem" has a higher score even though there are far fewer concepts that link to both "Tour Carpe Diem" and "Paris" than concepts that link to both "Eiffel Tower" and "Paris". To correct this, we should give a small boost to the relatedness score of two concepts that have many common inlinks.

[2]The out-neighbours of c: concepts to which c has a link.

The *shared inlink boost factor* (component $\mathcal{C}_2$), defined as

$$\mathcal{C}_2(x, y) = \log f(x, y)$$

is designed to give a boost to the relatedness score of two concepts that have many common inlinks. Here we use the logarithm because we do not want our measure to be linearly dependent on $f(x, y)$. In fact, this would result in stating that two concepts are related if and only if they are linked to by a large number of common concepts (as in the case of WLM). Although this is true when considering semantic relatedness, it does not always hold when considering spatial relatedness. For example, there are more concepts in Wikipedia that link to both "Paris" and "Rome" than concepts that link to both "Paris" and "Eiffel Tower", but "Eiffel Tower" is more spatially related to "Paris" than "Rome", as the former is one of the most important landmarks.

Graph Distance. Our third component $\mathcal{C}_3$ is based on the *graph distance* between two concepts in the Wikipedia graph. This component is based on the observation that the relatedness of two concepts x and y decreases with the graph distance $d(x, y)$. We have

$$\mathcal{C}_3(x, y) = \frac{1}{d(x, y)}$$

Note that when computing the distance between two concepts, we do not take into account the directions of the edges.

Combining the components. We finally combine the three components to obtain the formula of SYNRANK:

$$\text{SYNRANK}(x, y) = \mathcal{C}_1 \cdot \mathcal{C}_2 \cdot \mathcal{C}_3 = \frac{N \cdot f(x, y) \log f(x, y)}{f(x) \cdot f(y) \cdot d(x, y)}$$

$\mathcal{C}_2$ is multiplied to $\mathcal{C}_1$ in order to boost the score assigned by $\mathcal{C}_1$, which is based on PMI, of concepts that share many inlinks. From our experience, we do not obtain the same boost by summing the two components. Finally, we multiply by $\mathcal{C}_3$ as we noticed that the relatedness of two concepts rapidly decreases with their distance in the Wikipedia graph.

4.2 Using the Category Network

In this section we describe an approach, that we call CATEGORY, which creates the local lexicon of a location by using the Wikipedia *category network* C. According to the definition that can be found in Wikipedia itself, categories are meant to "help you to browse articles organized by concepts". This implies that each category corresponds to a concept and includes all articles that are relevant to it. For instance, the category named "Category:Washington, D.C." collects all the Wikipedia articles (and therefore concepts) related to concept "Washington, D.C.". It is easy to find the category that corresponds to a concept in Wikipedia, as it normally shares the same label as the concept or one of its redirects. Categories with the same name as the corresponding concepts are called *eponymous categories* in Wikipedia jargon. For any concept r, we denote its eponymous category as $EC(r)$. In the English Wikipedia, there are up to 100,000 eponymous categories. As a result, the extraction of local lexicons from Wikipedia categories is limited to those concepts that have corresponding eponymous categories.

Unfortunately the extraction of local lexicons from categories is far from being simple, and this is due to the structure of the category network in Wikipedia, which may lead to two problems that we term *concept drift* and *category incompleteness*, as defined below.

4.2.1 Concept drift

Let r be the root concept. Some of the concepts related to r are included in the category $EC(r)$ and the rest are spread through the subcategories of $EC(r)$. Therefore, the extraction of the local lexicon of r requires a visit of the subgraph of C rooted at $EC(r)$, which is induced by $EC(r)$ and all of its descendants. C is not a tree, which means that a descendant of $EC(r)$ may also belong to the subgraph rooted at a category other than $EC(r)$ that is not an ancestor of $EC(r)$. In other words, we observe a phenomenon that we call *concept drift*, where a category belonging to the subgraph rooted at $EC(r)$ might contain concepts not related to r. For instance, the subgraph rooted at the category labeled "Category:Washington, D.C." includes categories such as "Category: Executive Office of the President of the United States" and "Category:Recipients of the Langley Medal", whose concepts are loosely or not spatially related to "Washington, D.C.". In general, we remark that the more distant a category is from $EC(r)$, the less likely is it to include concepts related to r, where *distance* is defined as the length of the shortest path from $EC(r)$ to the category. However, this is only a trend and it is not clear how to threshold the distance from $EC(r)$ to discriminate in favor to the categories that include concepts related to r from those that do not.

4.2.2 Category incompleteness

It is fair to assume that the concepts related to r are primarily those in descendants of $EC(r)$, as category $EC(r)$, following the Wikipedia guidelines, is intended to collect such concepts. However, the parent categories of $EC(r)$ may also contain concepts related to r, as they may represent concepts more general than r, of which r is a particular aspect. For example, one of the parent categories of "New York City" is "Category:New York metropolitan area", and virtually every concept in it should be considered as spatially related to "New York City", because they correspond to locations that are geographically proximate to "New York City".

The algorithm that extracts the local lexicon of a concept from the categories, which we call Algorithm 1, takes as its input the root concept r, its eponymous *root category* $EC(r)$, which we recall has the same label as r, the Wikipedia graph W and the category network C. The output is the local lexicon $L(r)$ of the root concept. The first step consists of exploring all parent categories of $EC(r)$ and checking if they contain concepts related to r (lines 2–6). All the concepts in a parent category γ of $EC(r)$ are added to $L(r)$ if 50% or more of them link to r in the Wikipedia graph. In the second step, all concepts belonging to $EC(r)$ are added to $L(r)$ (line 7).

The recursive procedure VISIT is applied to explore the subgraph of C rooted at $E(r)$ with a DFS visit (lines 9–11). Line 8 initializes a set R which is used by VISIT to detect concept drift. Finally, each concept $c \in L(r)$ is assigned a score $1/d$, where d is the distance in C from $EC(r)$ of the category containing c (line 14).

The detection of concept drift lies at the heart of VISIT (Algorithm 2). Our solution consists of checking what we term the *admissibility* of each category γ (lines 2–7). If γ is found to be admissible, then the concepts included in γ are added to $L(r)$ and its subcategories are visited (lines 8–12). In order to decide whether a category γ is admissible, we use a list of what we term the *representative concepts* R, which best describe γ and its ancestors that have been visited already. Items in R are added as the DFS visit progresses through the subgraph rooted at $EC(r)$. When the visit starts at $EC(r)$, concept r is added to R (line 8 of Algorithm 1). When visiting a category γ other than $EC(r)$, any concept corresponding to γ (i.e., having the same label as γ) is added to R, along with all concepts corresponding to the parent categories of γ (line 2 of Algorithm 2). The addition to R of the parent categories

Algorithm 1 The CATEGORY algorithm

procedure CATEGORY($r, EC(r), W, C, L(r)$)
  for all parent γ of $EC(r)$ **do**
    if $\geq 50\%$ of the concepts in γ link to r **then**
      add all concepts included in γ to $L(r)$
    end if
  end for
  add r and all concepts included in $EC(r)$ to $L(r)$
  $R \leftarrow \{r\}$
  for all children γ of $EC(r)$ **do**
    VISIT($r, \gamma, W, C, L(r), R$)
  end for
  for all $c \in L(r)$ **do**
    d: distance in C of the category including c from $EC(r)$
    assign to c score $\frac{1}{d}$
  end for
end procedure

Algorithm 2 Visit the subgraph of C rooted at γ

procedure VISIT($r, \gamma, W, C, L(r), R$)
  Add to R all concepts corresponding to γ and γ's parents
  Extract $W(R, \gamma)$
  Sort nodes in $W(R, \gamma)$ by decreasing indegree
  n: number of nodes in $W(R, \gamma)$
  T_{max}: set of nodes of $W(R, \gamma)$ with highest indegree.
  if $r \in T_{max} \vee$ INDEGREE(r) $\geq \frac{n}{2} \vee \exists c \in T_{max}$ that links to r **then**
    add all concepts included in γ to $L(r)$
    for all children δ of γ **do**
      VISIT(r, $\delta, W, C, L(r), R$)
    end for
  end if
  remove from R all concepts added while visiting γ.
end procedure

of γ is the key to detecting concept drift. Indeed, since the category network is not a tree, the ancestors of γ are not limited to $EC(r)$ and the ancestors of $EC(r)$. Therefore, γ may have representative concepts that are unrelated to r (and therefore γ is not admissible). Using the "Category:Washington, D.C." example in the previous section, we see that most of the representative concepts of the parent categories of "Category:Recipients of the Langley Medal" have little to do with "Category:Washington, D.C.".

We now explain how we determine that the representative concepts of γ have "nothing to do" with γ. Recalling that concepts are nodes in the Wikipedia graph, we extract the subgraph $W(R, \gamma)$ of W induced by all concepts in R and all concepts in γ, and then sort the concepts in R by their indegree in $W(R, \gamma)$. If the root concept r belongs to the set T_{max} of the concepts with the highest indegree in $W(R, \gamma)$ then category γ is admissible. However, it would be too restrictive to consider γ admissible only in this case. In fact, γ is still a category of the subgraph rooted at $EC(r)$, which means that γ must be regarded as including concepts related to r, unless strong evidence to the contrary is found. Therefore, if $r \notin T_{max}$ but its indegree is still high, then γ must be considered as admissible. Finally, we observed that if a concept in T_{max} links to r, then γ is likely to contain concepts related to r. In short, γ is admissible if either $r \in T_{max}$, or 50% or more of the concepts included in γ link to r, or there is at least one concept in T_{max} that links to r.

Table 2: Local lexicons for "Paris"

JACC	JACCOPT	GREEN
France	France	France
London	Champs-Élysées	Paris-Gare de Lyon
Berlin	Montparnasse	Senate of France
Rome	5th arrondissement of Paris	Gare Montparnasse
Vienna	French Revolution	Transilien
Brussels	École Polytechnique	Optile
New York City	Axe historique	Voguéo
Italy	Île-de-France (region)	Gare du Nord
Belgium	Bastille Day	Corail (train)
French language	Hauts-de-Seine	Bondy

CATEGORY	ESA
Crépy-en-Valoiss	Paris Saint-Germain F.C.
Île-de-France	Paris FC
Converteam	Panthéon, Paris
A1 autoroute (France)	Basilique du Sacré-Coeur, Paris
Economy of Paris	RCF Paris
Île de la Jatte	Collège Stanislas de Paris
Île-de-France (province)	Paris Metro
Bernard de Lattre de Tassigny	Hôtel Ritz Paris
Émile Blanchard	Paris Olympia
Paul Gauguin	Economy of Paris

WLM	SYNRANK
France	Kilometre Zero
Haussmann's renovation of Paris	List of museums in Paris
Paris districts	Champs-Élysées
Champs-Élysées	La Défense
Musée du Louvre	Latin Quarter, Paris
Latin Quarter, Paris	RER C
History of Paris	Bois de Boulogne
Tuileries Palace	16th arrondissement of Paris
Lyon	Rive Droite
La Défense	Arrondissements of Paris

4.3 Examples

Table 2 shows a comparison of the top 10 concepts in the local lexicons of "Paris" obtained by using different relatedness measures, and illustrates their differences. The concepts in all local lexicons appear to be more or less semantically related to "Paris"; however, not all the measures are able to capture spatial relatedness. The local lexicon obtained with JACC includes concepts that are not spatially related to "Paris". While "London", "Berlin", "Rome", "Vienna", and "Brussels" can be considered as semantically related to "Paris", as they are all important European capitals, we expect to find a set of 10 concepts which characterize some aspect of "Paris", such as some of its landmarks or neighborhoods.

The local lexicon generated by JACCOPT contains concepts more directly spatially related to "Paris", including "Champs-Élysées", "Montparnasse", and "5th arrondissement of Paris". However, this local lexicon still contains concepts, such as "Bastille Day" and "French Revolution", that, even if they have some relatedness to "Paris", they should not appear in the top 10, as they are not just related to "Paris", but mainly to the more general concept "France".

The local lexicon created with CATEGORY is a bit peculiar, because of the way CATEGORY scores the concepts based on the distance of their category from the eponymous category of "Paris". Therefore, the concepts shown in the top 10 all belong to some cat-

egory that has distance 1 from the eponymous category of "Paris" and are all related to "Paris". In general, the CATEGORY measure, as we will show in Section 5, creates local lexicons with excellent quality. However, the concepts that have an eponymous category in Wikipedia only comprise 2% of Wikipedia, which severely limits this method's applicability.

The local lexicons obtained with GREEN and WLM have good quality but also contain concepts, such as "Optile" and "Lyon", that do not specifically imply "Paris", but respectively "Île-de-France", the region including "Paris", and "France".

Finally, local lexicons computed by SYNRANK and ESA have the best quality, as they only include concepts that are spatially related to PARIS. Being a text-based measure, ESA tends to assign higher scores to concepts whose label contains the word "Paris".

5. EVALUATION

In this section we describe the experiments that we conducted to evaluate JACC, JACCOPT, CATEGORY and SYNRANK, and to compare them to other existing relatedness measures. In particular, we compared them with GREEN [18] and WLM [17], which appear to be the best existing graph-based measures, and ESA [6], which is the best existing text-based measure.

One practical difficulty encountered in performing our experiments was dealing with large graph sizes. The Wikipedia graph contains millions of nodes and links, which cannot all reside in memory. To cope with these large graphs, we used the Large Sparse Graph Library[3], which was devised to handle large sparse graphs efficiently using techniques based on memory-mapped files. All experiments were executed on a Dell computer with 2 processors at 3GHz, with 8GB of memory, and running under Ubuntu.

We organize our evaluation in two parts. First, we evaluate the ability of the measures to capture the spatial relatedness (Subsection 5.1). Next, we show that the measures are also generalizable to compute semantic relatedness of Wikipedia articles (Subsection 5.2).

5.1 Spatial Relatedness

In order to evaluate the ability of a measure to create local lexicons, we select a subset of root concepts from Wikipedia and for each of them we create a ranked list of concepts sorted in descending order of their spatial relatedness score to the root concept. For each list, we select the top k concepts, and we check how many of them are actually spatially related to the corresponding root concept (which is referred to as *precision*); we also check how many of the Wikipedia concepts that are actually spatially related to the root concept are in the top k (*recall*).

Since here we consider the notion of spatial relatedness, we can use some properties of the Wikipedia in order to devise a semi-automatic procedure that determines whether two concepts are spatially related and asks humans to chime in only when the automatic procedure fails to make a decision. This allows us to make an evaluation on a high number of concepts and to study how the precision varies while increasing k, and further, what is a good value for k (up to 2000).

If we were to evaluate only precision and both the number of selected root concepts and k are small, then the ranked lists can be checked manually by humans, which is the approach commonly adopted in literature. Unfortunately, using human judges for evaluation is a huge manual effort, which in turn imposes practical limits on the number of concepts used in the evaluation. For example, Ollivier and Senellart [18] evaluated GREEN using just 7 concepts, with $k = 20$. However, an evaluation on such a small number of concepts is inevitably overly biased by the choice of concepts, and therefore is not reliable. Moreover, for concepts with high degree, we expect a relatedness measure to be able to accurately select many related concepts, not just 20.

Our automated evaluation procedure is based on the observation that Wikipedia provides some information that can be used to decide whether two concepts are spatially related or not. The primary information used in our evaluation consists of the spatial coordinate values attached to spatial concepts, such as "Paris". We also notice that there are *implicitly spatial* concepts that are not associated with spatial coordinate values, yet imply a location or a set of locations, termed a *spatial focus*. For example, "François Hollande" implies "France", as he is its president, while "Anne Hidalgo" implies "Paris", as she is its mayor. Therefore, we also associate spatial coordinate values to these implicitly spatial concepts, when this is possible. To do so, we use the procedure described in our previous work [21], which, given a non-spatial concept c, computes its spatial focus by clustering the spatial concepts in c's neighborhood, and computing the centroid of the most populous cluster, subject to a maximum distance radius.

In summary, two concepts are considered to be spatially related when they are spatially proximate, that is the distance between them is less than a given distance δ, whose value will be discussed shortly. This procedure allows us to load from Wikipedia a ground truth set of concepts $G(c)$ spatially related to a given concept c. If, while evaluating the top k concepts in a ranked list, a concept is found that does not belong to $G(c)$, then the concept is proposed to a human, who judges whether it is spatially related to c or not.

In our dataset of 1,200 spatial concepts, we found that each ranked list of k concepts with $k = 2000$ has on average only 130 concepts that need to be manually checked, which is a small and manageable number.

To test accuracy of the relatedness measures, we sampled 1,200 root concepts from Wikipedia in the following way. In order to evaluate whether the measures are sensitive to the degree of root concepts, we selected 400 concepts with low degree (300–1,000), 400 with high degree (1,000–5,000), and 400 with very high degree (over 5,000), termed the *small*, *medium*, and *large* datasets, respectively.

To create the datasets, we selected mostly cities, but also included nations and US states. For cities we set $\delta = 100$, as suggested by [15]. However, for the other two cases, we must slightly adapt our procedure to collect the ground truth. In Wikipedia, the spatial coordinate values of a nation or state usually correspond to the spatial coordinate values of its capital. Thus, if we select in the ground truth the concepts that are within δ miles of the capital, we create a set of concepts related to the capital, and not to the nation/state. Instead, to create the ground truth for nation and state concepts, we select those concepts that are contained in the nation/state.

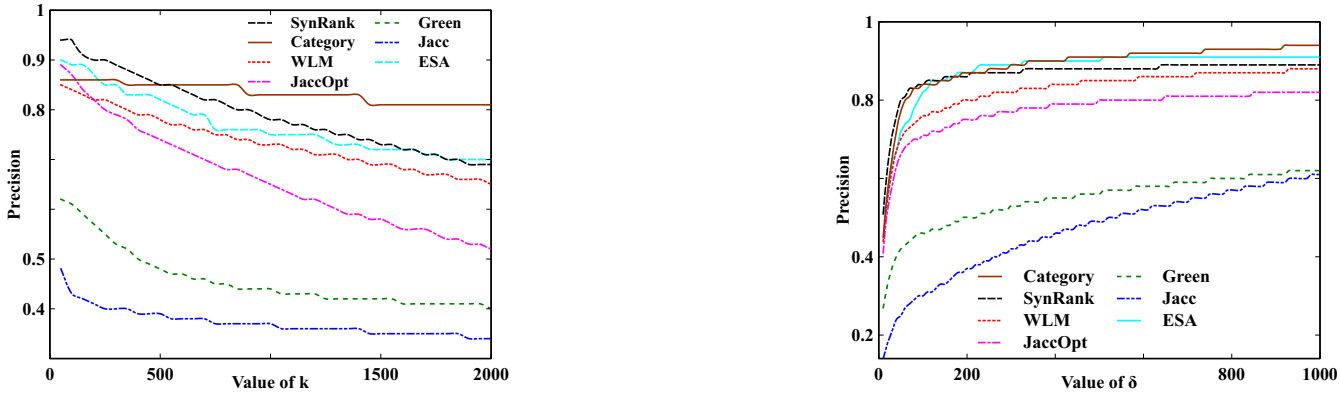
We now show how each measure described in this paper performs separately on each set of concepts. For each root concept c, each measure is used to create a ranked list $L(c)$ of k concepts. Precision $P(c)$ and recall $R(c)$ are computed as follows:

$$P(c) = \frac{|G(c)| \cap |L(c)|}{|L(c)|} \quad R(c) = \frac{|G(c)| \cap |L(c)|}{|G(c)|}$$

where $G(c)$ is the ground truth of c obtained with the procedure described earlier.

Figure 1 presents performance results in terms of precision, as measured over the *large* dataset, where precision is averaged over all concepts in the dataset. Figure 1a shows how the precision

[3] http://pierre.senellart.com/software/lsg



(a) Varying k

(b) Varying δ

Figure 1: Precision of each relatedness measure against k and δ

changes while varying k. As expected, with larger k, the precision decreases, which occurs when using all the described relatedness measures, including WLM and GREEN.

Overall, CATEGORY performs best, with a precision that is greater than 0.8, even for large values of k. This is not surprising, as CATEGORY retrieves the concepts related to a root concept r from its eponymous category $EC(r)$, which collects concepts that are explicitly related to r, along with some noise, as explained in Section 4.2. However, CATEGORY can only be used with root concepts that have an eponymous category, and as noted earlier, only 2% of the Wikipedia concepts in the English version have an eponymous category. On the other hand, SYNRANK is more general, as it can be applied to any Wikipedia concept. The precision of SYNRANK is above 0.9 for small values of k (50, 100) and is greater than 0.8, even with large values of k. As expected, SYNRANK performs better than WLM, which is due to the fact that the latter has been devised to capture a more general notion of relatedness between concepts. As a result, WLM considers that "Paris" and "Lyon" are strongly related because they are both major French cities, although they cannot be considered as spatially related.

As for ESA, there are several publicly available implementations; we chose ESALib [4], which is linked from the home page of the authors of ESA, because it provides a fast installation and configuration. ESA achieves a high precision, though slightly lower than SYNRANK. We note that ESA is considerably slower than SYNRANK and the other graph-based measures, but we are not sure that this is a problem of the implementation that we chose, or if it inherent to ESA.

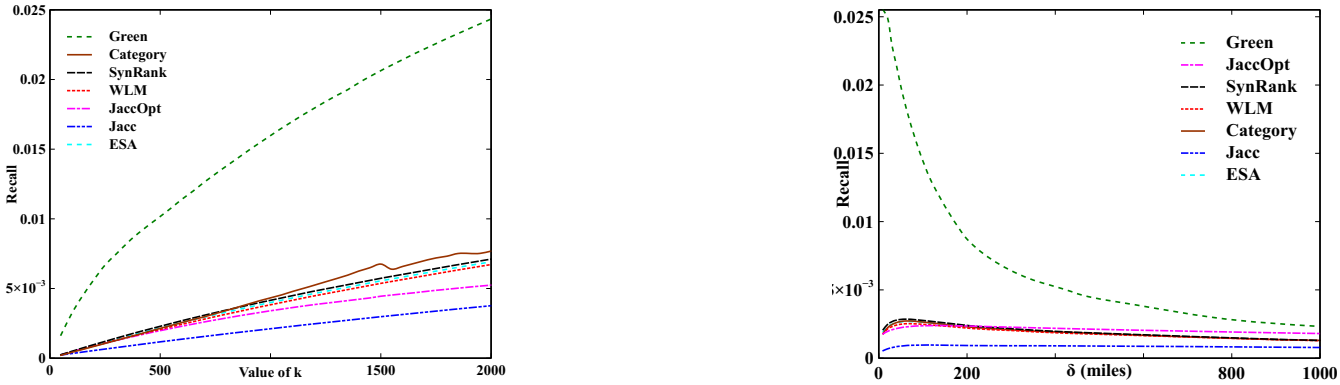
The low precision of GREEN is likely due to the fact that it gives high scores to concepts that share many common inlinks with the root concept, which, as explained in Section 4.1.2, are not necessarily spatially related to the root concept. This also explains the low precision of JACC which assigns high scores to concepts that share many inlinks and outlinks with the root concept. JACCOPT has been specifically obtained from JACC in order to capture spatial relatedness; not surprisingly, it achieves high precision, which is almost comparable to SYNRANK for small values of k. However, note that the precision of JACCOPT drops rapidly with increasing k. This is probably due to the fact that many Wikipedia articles lack an infobox, and most of them are *stubs*, which have just an introduction and nothing more. Therefore, in many cases, the heuristics of JACCOPT fail in selecting the important links of a concept.

Figure 1b shows how the methods' precision varies with δ. Here, k is fixed at 500. While this choice is somewhat arbitrary, we can see from Figure 1a that $k = 500$ guarantees high values of precision, at least for the best measures. For low values of δ (e.g., 10 miles), the ground truth is likely to contain only few concepts, resulting in low precision, since most of the top 500 concepts are not in the ground truth. As we increase the value of δ, the ground truth size increases, as well as the probability that one of the top 500 concepts is in the ground truth. Figure 1b shows that SYNRANK and CATEGORY greatly outperform the other measures for small values of δ, indicating their greater ability to select related concepts that are spatially proximate to the root concept. For larger values of δ (e.g., 1000 miles), the precision of SYNRANK, WLM, CATEGORY and ESA are essentially equal.

Figure 2 shows the performance in terms of recall, measured over the *large* dataset, and as before, averaged over all concepts in the dataset. Recall values are very low for every relatedness measure, which is likely due to the fact that Wikipedia contains many concepts with very low degree (the "stubs"), that are penalized by all the relatedness measures. We explain the high recall of GREEN with respect to the other measures by its iterative nature. At each iteration GREEN adjusts the scores of each node, based on the scores computed at the previous iterations, with the result of boosting the scores of the stubs more than the other relatedness measures. However, the relatively high values of recall happen at the expense of low precision (Figure 1). Not surprisingly, when k increases, the recall does as well (Figure 2a), since more concepts are selected. On the other hand, the recall is high when a low value of δ is selected (Figure 2b), as the ground truth is small. Except for JACC, the values of the recall are comparable for all the other measures; in particular, SYNRANK achieves values that are comparable to those obtained with both ESA and WLM.

Recall is not the primary criterion to evaluate the quality of a local lexicon, which does not need to be exhaustive; as we pointed out in our previous work [21], the key to determining the spatial reader scope of a news source is to have local lexicons that include concepts that are truly spatially related to a location. Large local lexicons that contain concepts that are not spatially related to a location can actually cause the determination of the wrong spatial reader scope. If we want to draw a parallel, with Web searches, users will not want a search engine to retrieve all possible results related to their queries, but rather expect that the first few results

[4] http://ticcky.github.io/esalib/



(a) Varying k (b) Varying δ

Figure 2: Recall of each relatedness measure against k and δ.

returned are correct and truly relevant to their queries.

As a last remark, we omit the graphs of precision and recall for the *medium* and *small* datasets, as they exhibit the same relative behaviour as those in Figures 1 and 2. We only note that for all similarity measures, the values for precision and recall are somewhat lower than those obtained with the *large* dataset. Thus, the degree of a concept has an influence on the quality of the local lexicons created with all the similarity measures described in this paper. A possible explanation for this is that a concept with small degree is likely to have a small set of related concepts, which are harder to identify in the wealth of Wikipedia concepts.

5.2 Semantic Relatedness

Since the determination of the semantic relatedness of two words or phrases is a well-studied problem, we can rely on existing datasets for our evaluation. Such datasets contain a list of word pairs along with a semantic relatedness score that is assigned by human evaluators. Since humans are able to determine whether two texts are semantically related, these datasets can be used as a gold standard. The goal of the evaluation is therefore to assess the extent to which the semantic relatedness scores assigned by a measure are in agreement with those assigned by humans. A common coefficient used for this purpose is the *Spearman rank-order correlation* ρ, which takes on values between -1 and +1; the higher the ρ, the better the agreement between the two sets of scores.

Two datasets are commonly used to evaluate semantic relatedness: the Wordsim-353 collection [4] and the Rubenstein-Goodenough's (RG) [22], which include 353 and 65 word pairs respectively. Both datasets are intended for evaluating the semantic relatedness between generic words. As a result, there may be pairs where either word does not have a corresponding article in Wikipedia (e.g. "loss", in the sense of negative difference between retail price and cost production) or is ambiguous, meaning that it may correspond to multiple articles (e.g. "keyboard", that may correspond to "Keyboard (computing)" or "Musical keyboard)"). If one pair contains one word that has no corresponding article, then we inevitably need to remove it. The same goes for the pairs where both words are ambiguous. If only one word in the pair is ambiguous, then we can still determine its corresponding article by looking at the other word, which may give some insights as to the correct sense of the ambiguous word. For instance, in the case of the word pair "computer, keyboard", we can select the article "Keyboard (computing)" as the one corresponding to the word "keyboard". As a result, we keep 202 word pairs in Wordsim-353 and 37 in RG.

Table 3: Evaluation of the semantic relatedness determination.

	WordSim-353		Rub.-Good.	
	ρ	Exec. time	ρ	Exec. (sec)
SYNRANK	**0.77**	0.42	0.75	0.09
WLM	0.73	0.51	0.70	0.10
JACCOPT	0.65	0.13	0.53	0.02
JACC	0.70	0.93	0.66	0.17
GREEN (1 iteration)	0.54	73.73	0.59	13.60
GREEN (2 iterations)	0.76	154.21	0.75	27.9
GREEN (3 iterations)	**0.77**	1202.94	**0.76**	205.9
GREEN (4 iterations)	**0.77**	3390.01	0.74	605.79
GREEN (5 iterations)	**0.77**	5589.37	0.74	1006.76
ESA	0.73	29	0.71	12

The results of the evaluation are given in Table 3. The table shows the value of ρ (the Spearman's rank order coefficient) and the execution times in seconds for each measure and dataset under evaluation. We did not evaluate CATEGORY because only a few articles in the two datasets have corresponding eponymous categories, which are necessary to use the measure. SYNRANK proves to be the best, as it obtains high correlation with the human evaluations on both datasets while being significantly faster than the other measures, except for JACCOPT which, however, achieves a significantly lower correlation. Good results are also obtained with WLM and JACC. The scores computed by ESA and GREEN obtain a high correlation with those assigned by humans, but they are much slower than SYNRANK. In particular GREEN needs at least two iterations (and further iterations do not improve its results).

In conclusion, SYNRANK seems to be a measure that is also able to capture a more general notion of semantic relatedness, although we devised it to compute the spatial relatedness of a concept to a location,

6. CONCLUDING REMARKS

In this paper we focused on the problem of determining the spatial relatedness of concepts (e.g., people, historical events, food specialties, movies, landmarks) to geographic locations (mostly, cities and countries). The interest in this problem stems from our previous work [21], in which we showed that the knowledge of

the spatial reader scope of a news source, that is the geographical location for which the content of the source has been primarily produced, is a key to disambiguating toponyms in a news article. The determination of the spatial reader scope of a news source requires the knowledge of a *local lexicon* of several geographic locations, that is a set of concepts that are spatially related to those locations.

In our previous work, we proposed a similarity measure that computes the spatial relatedness of a concept to a location by using the spatial coordinates assigned by Wikipedia to articles that describe spatial concepts (such as cities, landmarks). Although that approach works fine, the resulting local lexicons do not include non-spatial concepts (such as people, food specialties, historical events, movies) that are key components of the identity of a geographic location.

In this paper, we explored a set of *graph-based* similarity measures to determine a local lexicon of a location from Wikipedia without using any spatial clue, based on the observation that the spatial relatedness of a concept to a location is hidden in the Wikipedia link structure. Our evaluation on the local lexicons of 1,200 locations indicates that our observation is well-founded.

As we stressed before, recall is not the primary criterion to evaluate the quality of a local lexicon (precision is). Nonetheless, we intend to improve the poor recall achieved by all measures. Although we did not determine the precise reasons that negatively affect the recall, we suspect that the presence of many nodes with very low degree (referred to as "stubs" in the Wikipedia jargon) is one possible explanation. Indeed, a visual inspection of the local lexicons that we evaluated confirms that the occurrence of stubs is very low. How to modify the measures to boost the similarity score of stubs is part of our immediate future research directions. Another interesting evolution of our work would be to take into account the multilingual nature of Wikipedia.

Also, the semi-automatic nature of the evaluation raises some questions as to its reliability, especially in the case of the implicitly spatial concepts. As a future work we will include more labelled data (e.g. by using Mechanical Turk) to better assess the precision and the recall of the different measures.

An interesting survey presented by [9] shows that the coverage of a concept greatly varies across different language editions of Wikipedia, and so does the relatedness score between two concepts computed with the ESA similarity measure. That is, concepts that are described accurately and in detail in one language edition may be only briefly mentioned, or may not even exist, in another language version. Spatial concepts are examples of concepts with differing coverage in Wikipedia language editions. Therefore it would be interesting to understand how to leverage these differences in coverage to create local lexicons of locations.

Acknowledgment

This work was supported in part by the National Science Foundation under Grants IIS-10-18475, IIS-12-19023, and IIS-13-20791.

7. REFERENCES

[1] R. L. Cilibrasi and P. M. B. Vitanyi. The Google similarity distance. *IEEE TKDE*, 19(3):370–383, 2007.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[3] C. Esperança and H. Samet. Experience with SAND/Tcl: a scripting tool for spatial databases. *Journal of Vis. Lang. and Comput,*, 13(2):229–255, 2002.

[4] L. Finkelstein, E. Gabrilovych, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. *ACM TOIS*, 20(1):116–131, 2002.

[5] B. C. Fruin, H. Samet, and J. Sankaranarayanan. Tweetphoto: photos from news tweets. In *GIS*, pages 582–585, Redondo Beach, CA, 2012.

[6] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611, Hyderabad, India, 2007.

[7] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[8] N. Gramsky and H. Samet. Seeder finder - identifying additional needles in the Twitter haystack. In *LBSN*, pages 44–53, Orlando, FL, 2013.

[9] B. Hecht and D. Gergle. On the "localness" of user-generated content. In *CSCW*, pages 229–232, Savannah, GA, 2010.

[10] P. Jaccard. Etude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[11] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using Twitter. In *LBSN*, pages 25–32, Chicago, 2011.

[12] M. D. Lieberman and H. Samet. Multifaceted toponym recognition for streaming news. In *SIGIR*, pages 843–852, Beijing, China, 2011.

[13] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *SIGIR*, pages 731–740, Portland, OR, 2012.

[14] M. D. Lieberman and H. Samet. Supporting rapid processing and interactive map-based exploration of streaming news. In *GIR*, pages 179–188, Redondo Beach, CA, 2012.

[15] M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE*, pages 201–212, Long Beach, CA, 2010.

[16] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. STEWARD: architecture of a spatio-textual search engine. In *GIS*, pages 186–193, Seattle, WA, 2007.

[17] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *WikiAI'08: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 25–30, Chicago, 2008.

[18] Y. Ollivier and P. Senellart. Finding related pages using Green measures: An illustration with Wikipedia. In *AAAI*, pages 1427–1433, Vancouver, Canada, 2007.

[19] P. Pantel and D. Lin. Discovering word senses from text. In *KDD*, pages 613–619, Edmonton, Canada, 2002.

[20] S. P. Ponzetto and M. Strube. Knowledge derived from Wikipedia for computing semantic relatedness. *JAIR*, 30(1):181–212, 2007.

[21] G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman. Determining the spatial reader scopes of news sources using local lexicons. In *GIS*, pages 43–52, San Jose, CA, 2010.

[22] H. Rubenstein and J. B. Goodenough. Contextual correlates of synonymy. *CACM*, 8(10):627–633, 1965.

[23] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and J. Sankaranarayanan. PhotoStand: A map query interface for a database of news photos. *PVLDB*, 6(12):1350–1353, 2013.

[24] H. Samet, M. D. Adelfio, B. C. Fruin, M. D. Lieberman, and B. E. Teitler. Porting a web-based mapping application to a smartphone app. *GIS*, pages 525–528, Chicago, 2011.

[25] H. Samet, H. Alborzi, F. Brabec, C. Esperança, G. R. Hjaltason, F. Morgan, and E. Tanin. Use of the SAND spatial browser for digital government applications. *CACM*, 46(1):63–66, 2003.

[26] H. Samet, J. Sankaranarayanan, M. D. Lieberman, M. D. Adelfio, B. C. Fruin, J. M. Lotkowski, D. Panozzo, J. Sperling, and B. E. Teitler. Reading news with maps by exploiting spatial synonyms. *CACM*, 57(10), 2014.

[27] H. Samet, B. E. Teitler, M. D. Adelfio, and M. D. Lieberman. Adapting a map query interface for a gesturing touch screen interface. In *WWW (Companion Volume)*, pages 257–260, Hyderabad, India, 2011.

[28] J. Sankaranarayanan and H. Samet. Images in news. In *ICPR*, pages 3240–3243, Istanbul, Turkey, 2010.

[29] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in tweets. In *GIS*, pages 42–51, Seattle, WA, 2009.

[30] B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS*, pages 144–153, Irvine, CA, 2008.

[31] D. Turdakov and P. Velikhov. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In *SYRCoDIS08: Proceedings of the 5th Spring Young Researchers Colloquium on Databases and Information Systems*, Saint Petersburg, Russia, 2008.

[32] S. Wubben and A. van den Bosch. A semantic relatedness metric based on free link structure. In *ICCS*, pages 355–358, Tilburg, The Netherlands, 2009.