Finding Related Pages Using Green Measures: An Illustration with Wikipedia

Yann Ollivier, Pierre Senellart

Finding Related Pages Using Green Measures: An Illustration with Wikipedia

Yann Ollivier
CNRS & École normale supérieure de Lyon
46 allée d'Italie
69364 Lyon Cedex 07, FRANCE
yann.ollivier@normalesup.org

Pierre Senellart
INRIA Futurs & Université Paris-Sud
4 rue Jacques Monod
91893 Orsay Cedex, FRANCE
pierre@senellart.com

Abstract

We introduce a new method for finding nodes semantically related to a given node in a hyperlinked graph: the Green method, based on a classical Markov chain tool. It is generic, adjustment-free and easy to implement. We test it in the case of the hyperlink structure of the English version of Wikipedia, the on-line encyclopedia. We present an extensive comparative study of the performance of our method versus several other classical methods in the case of Wikipedia. The Green method is found to have both the best average results and the best robustness.

Introduction

The use of tools relying on graph structures for extracting semantic information in a hyperlinked environment (Kleinberg 1999) has had vast success, which led to a revolution in the search technology used on the World Wide Web (Brin & Page 1998). In the same spirit, we present here a novel application of a classical tool from Markov chain theory, the Green measures, to the extraction of semantically related nodes in a directed graph. Such a technique can help a user find additional content closely related to a node i and thus guide her in the exploration of a graph. Google and Google Scholar both allow users to search for similar nodes, respectively in the Web graph and in the graph of scientific publications. This could also be useful in the case of the graph of an on-line encyclopedia like Wikipedia, where articles are seen as nodes of the graph and hyperlinks as edges between nodes: users are often interested in looking for articles on related topics, for instance to deepen their understanding of some concept. Other interests of an automatic method for finding related articles can be for instance to add missing links between articles (Adafre & de Rijke 2005).

Our proposed method can be intuitively described as a PageRank (Brin & Page 1998) computation that continuously pours mass at node i. It is related to, but distinct from, so-called topic-sensitive PageRank (Haveliwala 2003) (see below). The method provides a measure for similarity of nodes and could serve as a definition for some kind of "conceptual neighborhood" around i.

In order to be able to have a somewhat objective measure of the performance of the Green method, we compared it to several more classical approaches for extracting related pages in a graph. All these methods have been implemented, and tested on the graph of the English version of Wikipedia; though, to preserve generality of the approach, we did not implement any Wikipedia-specific tricks to enhance performance. A user study has been performed which allows us to evaluate and compare each of these techniques.

Our contributions are thus: (i) a novel use of Green measures to extract semantic information in a graph (ii) an extensive comparative study of the performance of different methods for finding related articles in the Wikipedia context. Note that we implemented "pure" versions of the methods: it is certainly easy to devise Wikipedia-specific enhancements to the methods, but we refused to do so in order to keep the comparison general. Even so, performance of the Green method was very satisfying.

We first introduce Green measures. Then we present different methods for extracting related nodes in a graph, based on Green measures, on PageRank, and on classical Information Retrieval approaches. The results of the experiment carried out to compare these methods are described next. Finally, we discuss related work and perspectives.

Additional data about the content presented here (including source code and full evaluation results) is available on the companion website for this paper (Ollivier & Senellart 2007).

Green Measures

Notation for Markov Chains. We collect here some standard facts and notation about Markov chains (Norris 1997).

Let (p_{ij}) be the transition probabilities of a Markov chain on a finite set of states X. That is, each p_{ij} is a non-negative number representing the probability to jump from node $i \in X$ to node $j \in X$; in particular, for each i we have $\sum_j p_{ij} = 1$. That is, the p_{ij}'s form a stochastic matrix.

For example, if X is given as a directed graph, we can define the *simple random walk on* X by setting $p_{ij} = 0$ if there is no edge from i to j, and $p_{ij} = 1/d_i$ if there is an edge from i to j, where d_i is the number of edges originating from i (if multiple edges are allowed, this definition can be adapted accordingly). This remark is very important since it allows one to view any hyperlinked environment as a Markov chain and to use and/or adapt Markov chain techniques.

A row vector $\mu = (\mu_i) : X \to \mathbb{R}$ indexed by X will be called a *measure on* X (negative values are allowed). The *(total) mass* of μ is $\sum \mu_i$. If moreover $\mu_i \geqslant 0$ and $\sum \mu_i = 1$, the measure will be called a *probability measure*.

We define the *forward propagation operator* M as follows: for any measure $\mu = (\mu_i)$ on X, the measure μM is defined by $(\mu M)_j := \sum_i \mu_i p_{ij}$, that is, each node i sends a part p_{ij} of its mass to node j. This corresponds to multiplication on the right by the matrix $M = (p_{ij})$, hence the notation μM. Note that forward propagation preserves the total mass $\sum \mu_i$.

Henceforth, we suppose, in a standard manner, that the Markov chain is irreducible and aperiodic (Norris 1997). For the simple random walk on a graph, it amounts to the graph being strongly connected and the greatest common divisor of the lengths of all cycles being equal to 1.

Under all these assumptions, it is well-known that the Markov chain has a unique invariant probability measure ν, the *equilibrium measure*: that is, a unique measure ν with $\nu M = \nu$ and $\sum \nu_i = 1$. Moreover, for any measure μ such that $\sum \mu_i = 1$, the iterates μM^n converge to ν as $n \to \infty$. More precisely, the matrices M^n converge exponentially fast (in the number of iterations) to a matrix M^∞, which is of rank 1 and satisfies $M^\infty_{ij} = \nu_j$ for all i. The equilibrium measure ν can be thought of as a PageRank without random jumps on X (Brin & Page 1998).

Definition of Green Measures. Green functions were historically introduced in electrostatic theory as a means of computing the potential created by a charge distribution; they have later been used in a variety of problems from analysis or physics (Duffy 2001), and extended to discrete settings. The Green measure centered at i, as defined below, can really be thought of as the electric potential created on X by a unit charge placed at i (Kemeny, Snell, & Knapp 1966).

The *Green matrix* of a finite Markov chain is defined by

$$G := \sum_{t=0}^{\infty} (M^t - M^\infty)$$

where M^t is the t-th power of the matrix $M = (p_{ij})$, corresponding to t steps of the random walk. Since the M^t converge exponentially fast to M^∞, the series converges.

Now, for $i \in X$, let us define G_i, the *Green measure centered at* i, as the i-th row of the Green matrix G.

Let δ_i be the Dirac measure centered at i, that is, $\delta_{ij} := 1$ if $j = i$ and 0 otherwise. We have by definition $G_i = \delta_i G$. More explicitly, using that $M^\infty_{ij} = \nu_j$, we get

$$G_{ij} = \sum_{t=0}^{\infty} \left((\delta_i M^t)_j - \nu_j \right)$$

where $(\delta_i M^t)_j$ is of course the probability that the random walk starting at i is at j at time t. Since $\delta_i M^t$ is a probability measure and ν is as well, we see that for each i, G_i is a measure of total mass 0.

We now present other natural interpretations of the Green measures (in addition to electric potential).

PageRank with Source at i. The sum

$$G_i = \sum_{t=0}^{\infty} (\delta_i - \nu) M^t$$

is a fixed point of the operator

$$\mu \mapsto \mu M + (\delta_i - \nu)$$

This fixed point is thus the equilibrium measure of a random walk with a source term δ_i which constantly pours a mass 1 at i, and a uniform sink term $-\nu$ (to preserve total mass). This is what makes Green measures different from PageRank and focused around a node.

This shows how Green measures can be computed in practice: Start with the row vector $\mu = \delta_i - \nu$ and iterate $\mu \mapsto \mu M + (\delta_i - \nu)$ some number of times. This allows to compute the Green measure centered at i without computing the whole Green matrix.

Time Spent at a Node. Since the equilibrium measure is ν, the average time spent at any node $j \in X$ by the random walk between steps 0 and t behaves, in the long run, like $(t+1)\nu_j$, whatever the starting node was. Knowing the starting node i leads to a correction term, which is precisely the Green measure centered at i. More precisely:

$$G_{ij} = \lim_{t \to \infty} \left(T_{ij}(t) - (t+1)\nu_j \right)$$

where $T_{ij}(t)$ is the average number of times the random walk starting at i hits node j between steps 0 and t (included).

Relationship with Topic-Sensitive PageRank. Topic-sensitive PageRank is a method for answering keyword queries on the World Wide Web which biases the PageRank method by focusing the PageRank computation around some *seed* subset of pages (Haveliwala 2003). It proceeds as follows. First, a list of topics is fixed, for each of which a list of *seed* Web pages is determined by hand. Second, for each different topic, a modified Markov chain is used which consists in, at each step, either following an outlink with probability $1 - c$, or jumping back to a seed page with probability c. Third, when answering queries, these modified PageRank values are combined with weights depending on the frequency of query terms in the seed documents.

Green measures are somewhat related to the modified Markov chain used as the second ingredient of topic-sensitive PageRank. Namely, let i be a single node that we use as the seed. Then the matrix $\tau(c)$ whose entry $\tau_{ij}(c)$ is the value at node j of the topic-sensitive PageRank with seed i is easily seen to be

$$\tau(c) = \sum_{t=0}^{\infty} c(1-c)^t M^t$$

where M is the transition matrix of the original Markov chain, and as above c is the rate at which the random walk jumps back to the seed. When c tends to 0, of course topic-sensitivity is lost, and the series tends to the matrix M^∞ all

the rows of which are equal to the ordinary PageRank vector ν.

Now, we have:

$$\tau(c) - M^\infty = \left(\sum_{t=0}^{\infty} c(1-c)^t M^t \right) - M^\infty$$
$$= \sum_{t=0}^{\infty} c(1-c)^t \left(M^t - M^\infty \right)$$

thanks to the identity $\sum c(1-c)^t = 1$. The Green matrix is thus related to the topic-sensitive matrix τ as follows:

$$G_{ij} = \lim_{c \to 0} \frac{1}{c} \left(\tau_{ij}(c) - \nu_j \right)$$

Thus, yet another interpretation of Green measures is as a way to get rid of the tendency of topic-sensitive PageRank to reproduce the global PageRank, and to extract meaningful information from it in a canonical way, without an arbitrary choice of the coefficient c.

Description of the Methods

We now proceed to the definition of the five methods included in the evaluation: two Green-based methods and three more classical approaches.

The goal of each method is, given a node i in a graph (or in a Markov chain), to output an ordered list of nodes which are "most related" to i in some sense. All methods used here rely on scoring: given i, every node j is attributed a score $S^i(j)$. We then output the n nodes with highest score. Here we arbitrarily set $n = 20$, as we could not devise a natural and universal way to define a threshold.

Two Green-Based Methods

GREEN. The **GREEN** method relies directly on the Green measures described above. When looking for nodes similar to node i, compute the Green measure G_i centered at i. Now for each j, the value G_{ij} indicates how much node j is related to node i and can be taken as the score $S^i(j)$.

This score leads to satisfying results. However, nodes j with higher values of the equilibrium measure ν_j were slightly overrepresented. We found that performance was somewhat improved if an additional term favoring uncommon nodes j is introduced. Namely, we set

$$\boxed{S^i(j) := G_{ij} \log(1/\nu_j)}$$

The logarithmic term comes from information theory: $\log(1/\nu_j)$ is the quantity of information brought by the event "the random walk currently lies at node j", knowing that its prior probability is ν_j. This is very similar to the logarithmic term in the tf-idf formula used for **COSINE** below.

SYMGREEN. Since it mainly consists in following the Markov chain flow starting at node i, **GREEN** might miss nodes that point to i but are not pointed to by i, nodes which could be worth considering. The workaround is to symmetrize the Markov chain as follows: Given any Markov chain (p_{ij}) with stationary measure $\nu = (\nu_i)$, the *symmetrized Markov chain* is defined by

$$\tilde{p}_{ij} := \frac{1}{2} \left(p_{ij} + p_{ji} \frac{\nu_j}{\nu_i} \right)$$

which is still a stochastic matrix. This definition is designed so that the new Markov chain still has the same equilibrium measure ν. (Observe that simply forgetting edge orientation is not a proper way to symmetrize ν, since it will result in an invariant measure proportional to the degree of the node and ignore the actual values of the probabilities.)

This amounts to, at each step, tossing a coin and following the origin Markov chain either forward or backward (where the backward probabilities are given by $p_{ji}\nu_j/\nu_i$).

The Green measures $\tilde{G}_i$ for this new Markov chain $(\tilde{p}_{ij})$ can be defined in the same way, and as above the scores are given by

$$\boxed{S^i(j) := \tilde{G}_{ij} \log(1/\nu_j)}$$

PageRank-Based Methods

Arguably the most naive method for finding nodes related to a given node i is to look at nodes with high PageRank index in a neighborhood of i. Similar techniques are extensively used for finding *related pages* on the World Wide Web (Kleinberg 1999; Dean & Henzinger 1999). Here by PageRank we mean the equilibrium measure of the random walk, that is, we discard random jumps (we set Google's *damping factor* to 1). Indeed, random jumps tend to spread the equilibrium measure more uniformly on a graph, whereas the goal here is to focus around a given node.

We describe two ways of using the equilibrium measure to identify nodes related to a given node.

PAGERANKOFLINKS. The first method that springs to mind for identifying nodes related to i is to take the nodes pointed to by i and output them according to their PageRank.

Namely, let ν be the equilibrium measure of the random walk on the graph (or of the Markov chain). Let i be a node. The score of node j in the **PAGERANKOFLINKS** method is defined by

$$\boxed{S^i(j) := \begin{cases} \nu_j & \text{if} \quad p_{ij} > 0 \\ 0 & \text{if} \quad p_{ij} = 0 \end{cases}}$$

LOCALPAGERANK. Another PageRank-based method was implemented. It consists in, first, building a restricted graph centered at node i (namely, nodes obtained by following the links forwards, backwards, forwards-backwards and backwards-forwards), and then computing the equilibrium measure on this subgraph. The method outputs nodes of this subgraph, ordered according to this "local PageRank".

This method has an important flaw: As soon as the graph is highly connected, as is the case with Wikipedia, the neighborhood comprises a significant portion of the original graph. In such a case, the local equilibrium measure is very close to the global equilibrium measure, and so the results are not at all specific to i.

Due to its extremely poor results, this method was not included in the test. For example, on *Pierre de Fermat* the first 10 results in the output are *France*, *United States*, *United Kingdom*, *Germany*, *2005*, *2006*, *World War II*, *Italy*, *Europe* and *England*, showing no specific relationship to the base article but close to the global PageRank values.

Information Retrieval-Inspired Methods

Standard information retrieval methods can be applied when only a graph/Markov chain is available, provided one is able to define the "content" of a node i. It is natural to interpret the set of nodes pointed to by i as the content of i, and moreover the transition probabilities p_{ij} can be thought of as the frequency of occurrence of j in i.

We tested two such methods: a cosine method using a tf-idf weight, and a cocitation index method.

COSINE with tf-idf weight. Cosine computations first use some transformation to represent each node/document in the collection by a vector in $\mathbb{R}^n$ for some fixed n. The proximity of two such vectors can then be measured by their cosine as ordinary vectors in $\mathbb{R}^n$ (or their angle, which amounts to the same as far as ordering is concerned).

One very usual such vector representation for documents is given by the *term frequency/inverse document frequency (tf-idf) weight* (Salton & McGill 1984). In our setting, it is adapted as follows.

Given a Markov chain defined by (p_{ij}) on a set of N elements (e.g. the random walk on a graph), for each node i the tf-idf vector x^i associated with i is an N-dimensional vector defined by

$$(x^i)_j := p_{ij} \log (N/d_j)$$

where d_j is the number of nodes pointing to j.

COSINE is then very simple: when looking for nodes related to node i, the score of node j is defined by

$$\boxed{S^i(j) := \cos(x^i, x^j)}$$

where x^i and x^j are seen as vectors in $\mathbb{R}^N$. Here, as usual, $\cos(x, y) = \frac{\sum x_k y_k}{\sqrt{\sum x_k^2}\sqrt{\sum y_k^2}}$.

COCITATIONS. A standard and straightforward method to evaluate document similarity is the cocitation index: two documents are similar if there are many documents pointing to both of them. This method, which originated in bibliometrics, is well-known and widely used for similar problems, see for instance (Dean & Henzinger 1999) for an application to the Web graph.

In our context this simply reads as follows. When looking for nodes similar to a node i, the score of node j is given by

$$\boxed{S^i(j) := \#\{k,\, p_{ki} > 0 \text{ and } p_{kj} > 0\}}$$

Sometimes this method tends to favor nodes that have the same "type" as i rather than nodes semantically related to i but with a different nature. For example, when asked for pages related to *1989* (the year) in Wikipedia, the output is *1990*, *1991*... For the base article *Pierre de Fermat*, interestingly, it outputs several other great mathematicians.

Experimental Results

In this section, we describe the experiments carried out to evaluate the performance of the methods presented above, on the graph of the English version of Wikipedia.

Graph Extraction, Implementation. A September 25th, 2006 dump of the English Wikipedia was downloaded from the URL `http://download.wikimedia.org/`. It was parsed in order to build the corresponding directed graph: nodes are the Wikipedia pages, and edges represent the links between pages. Multiple links were kept as multiple edges. Redirections (alternate titles for the same entry) were resolved. The most common templates (*Main*, *See also*, *Further*, *Details*...) were expanded. Categories (special entries on Wikipedia which are used to group articles according to semantic proximity, such as *Living people*) were kept as standalone pages, just as they appear on Wikipedia.

The resulting graph has $1,606,896$ nodes and $38,896,462$ edges; there are $73,860$ strongly connected components, the largest one of which contains $1,531,989$ nodes. We restrict ourselves to this strongly connected subgraph, in order to ensure convergence of computation of the equilibrium measure and Green measures.

Implementation of the methods is mostly straightforward, but here are a few caveats: 1. Because of the large size of the graph, memory handling must be considered with care; a large sparse graph library, relying on memory-mapped files, has been developed for this purpose. 2. Most methods require prior knowledge of the equilibrium measure for the graph, which is therefore computed once with very high accuracy. 3. Rather than the Green matrix, we compute the Green measure centered at i using the characterization of Green measures as fixed point of the operator $\mu \mapsto \mu M + (\delta_i - \nu)$.

The computation time for **GREEN** is less than 10s per article on a 3GHz desktop PC; that of **SYMGREEN** is typically between 15s and 30s. The other methods range from a few seconds to three minutes (**COSINE**). Computation of **GREEN** is easily parallelizable; we estimate that computation of the full Green matrix would take less than two weeks on a 10 PC cluster, after which the answers are instantaneous.

Evaluation Methodology. We carried out a blind evaluation of the methods on 7 different articles, chosen for their diversity: (i) *Clique (graph theory)*: a very short, technical article. (ii) *Germany*: a very large article. (iii) *Hungarian language*: a medium-sized, quite technical article. (iv) *Pierre de Fermat*: a short biographical article. (v) *Star Wars*: a large article, with an important number of links. (vi) *Theory of relativity*: a short introductory article pointing to more specialized articles. (vii) *1989*: a very large article, containing all the important events of year 1989. It was unreasonable to expect our testers to evaluate more articles. In order to avoid any bias, we did not run the methods on these 7 articles before the evaluation procedure was launched.

People were asked to assign a mark between 0 and 10 (10 being the best) to the list of the first 20 results returned by

Table 1: Output of **GREEN** on the articles used for evaluation.

Clique (graph theory)	*Germany*	*Hungarian language*	*Pierre de Fermat*	*Star Wars*	*Theory of relativity*	*1989*
1. Clique (graph theory) 2. Graph (mathematics) 3. Graph theory 4. Category:Graph theory 5. NP-complete 6. Complement graph 7. Clique problem 8. Complete graph 9. Independent set 10. Maximum common subgraph isomorphism problem 11. Planar graph 12. Glossary of graph theory 13. Mathematics 14. Connectivity (graph theory) 15. Computer science 16. David S. Johnson 17. Independent set problem 18. Computational complexity theory 19. Set 20. Michael Garey	1. Germany 2. Berlin 3. German language 4. Christian Democratic Union (Germany) 5. Austria 6. Hamburg 7. German reunification 8. Social Democratic Party of Germany 9. German Empire 10. German Democratic Republic 11. Bavaria 12. Stuttgart 13. States of Germany 14. Munich 15. European Union 16. National Socialist German Workers Party 17. World War II 18. Jean Edward Smith 19. Soviet Union 20. Rhine	1. Hungarian language 2. Slovakia 3. Romania 4. Slovenia 5. Hungarian alphabet 6. Hungary 7. Croatia 8. Category:Hungarian language 9. Turkic languages 10. Finno-Ugric languages 11. Austria 12. Serbia 13. Uralic languages 14. Ukraine 15. Hungarian grammar (verbs) 16. German language 17. Hungarian grammar 18. Khanty language 19. Hungarian phonology 20. Finnish language	1. Pierre de Fermat 2. Toulouse 3. Fermat's Last Theorem 4. Diophantine equation 5. Fermat's little theorem 6. Fermat number 7. Grandes écoles 8. Blaise Pascal 9. France 10. Pseudoprime 11. Lagrange's four-square theorem 12. Number theory 13. Fermat polygonal number theorem 14. Holographic will 15. Diophantus 16. Euler's theorem 17. Pell's equation 18. Fermat's theorem on sums of two squares 19. Fermat's spiral 20. Fermat's factorization method	1. Star Wars 2. Dates in Star Wars 3. Palpatine 4. Jedi 5. Expanded Universe (Star Wars) 6. Star Wars Episode I: The Phantom Menace 7. Star Wars Episode IV: A New Hope 8. Obi-Wan Kenobi 9. Star Wars Episode III: Revenge of the Sith 10. Coruscant 11. Anakin Skywalker 12. Lando Calrissian 13. Luke Skywalker 14. Star Wars: Clone Wars 15. List of Star Wars books 16. George Lucas 17. Star Wars Episode II: Attack of the Clones 18. Splinter of the Mind's Eye 19. List of Star Wars comic books 20. The Force (Star Wars)	1. Theory of relativity 2. Special relativity 3. General relativity 4. Spacetime 5. Lorentz covariance 6. Albert Einstein 7. Principle of relativity 8. Electromagnetism 9. Lorentz transformation 10. Inertial frame of reference 11. Speed of light 12. Galilean transformation 13. Local symmetry 14. Category:Relativity 15. Galilean invariance 16. Gravitation 17. Global symmetry 18. Tensor 19. Maxwell's equations 20. Introduction to general relativity	1. 1989 2. Cold War 3. 1912 4. Tiananmen Square protests of 1989 5. Soviet Union 6. German Democratic Republic 7. George H. W. Bush 8. 1903 9. Communism 10. 1908 11. 1929 12. Ruhollah Khomeini 13. March 1 14. Czechoslovakia 15. June 4 16. The Satanic Verses (novel) 17. 1902 18. November 7 19. October 9 20. March 14
Mark: 7.6/10	Mark: 7.0/10	Mark: 6.2/10	Mark: 7.3/10	Mark: 7.4/10	Mark: 8.1/10	Mark: 5.4/10

each method on these articles, according to their relevance as "related articles" lists. Each evaluator was free to interpret the meaning of the phrase "related articles". The lists were unlabeled, randomly shuffled, and in a potentially different order for each article. The evaluators were allowed to skip articles they did not feel confident enough to vote on. There has been a total of 67 participants, which allows for reasonably good confidence intervals.

Performance of the Methods. Table 1 shows the output of **GREEN** on each evaluated article. Due to lack of space, we only present a portion of the outputs of the other methods in Table 2. The full output and detailed evaluation results can be found in (Ollivier & Senellart 2007).

The average marks given by the evaluators are presented in a radar chart on Figure 1. Each axis stands for the mark given for an article: from worst $(0/10)$ at the center to best $(10/10)$ at the periphery, while each row represents a method (cf. the legend). Table 3 gives global statistics about the performance of the methods.

Absolute marks should be taken with caution: it is probable that a human-designed list of related pages would not score very close to $10/10$, but maybe closer to $8/10$. Indeed, the evaluator-to-evaluator standard deviation for a given article is always between 1.5 and 2.0. For example, on *Theory of relativity*, **GREEN** gets $8.1/10$ though it was attributed a top $10/10$ mark by a significant number of evaluators, including several experts in this field.

GREEN presents the best overall performance. The difference between global scores of **GREEN** and of the best classical approach, **COSINE**, is 1.8, which is statistically significant. **GREEN** comes out first for all but two articles, where it is second with a hardly significant gap (0.4 in both cases). Moreover, **GREEN** is extremely robust: First, it has a low article-to-article standard deviation, and a look at Figure 1 shows that it never performs very badly. Second, there are very few irrelevant words in its output, as can be seen on Table 1; the high number of $10/10$ given to **GREEN** is perhaps a measure of this fact. Finally, some of the related articles proposed by **GREEN** are both highly semantically relevant and completely absent from the output of other methods: this is the case of *Finnish language* for *Hungarian language* (linguists now consider both languages closely related), and of *Tiananmen Square protests* or *The Satanic Verses* for *1989*.

SYMGREEN presents a profile similar to **GREEN** for both performance and robustness. Actually, though its overall mark is slightly less on the evaluated articles, on other articles we experimented with in an informal way, it seems more robust than **GREEN**. It might in fact be better adapted for other contexts, especially in less highly connected graphs.

COSINE performs best of the "classical" methods, but is clearly not as good as the Green-based ones. Both very good and very bad performance occur: compare for instance *Germany* and *Pierre de Fermat* in Table 2. Thus, this method is unstable, which is visible in its high article-to-article standard deviation. Moreover, even in the case when it performs well, as for *Germany*, completely irrelevant or anecdotal entries are proposed, like *Pleasure Victim* or *Hildesheimer Rabbinical Seminary*. Testing the methods informally on more articles confirmed this serious instability of **COSINE**.

COCITATIONS does not give very good results, but it is still interesting: more than *related* articles, it outputs lists of articles of the same *type*, giving for instance names of great mathematicians of the same period for *Pierre de Fermat*, languages for *Hungarian language* or years for *1989*.

PAGERANKOFLINKS is the worst of the methods tested (although **LOCALPAGERANK**, not formally tested here, is even worse). It basically outputs variations on the global

Table 2: Output of **SYMGREEN**, **COSINE**, **COCITATIONS**, and **PAGERANKOFLINKS** on sample articles.

SYMGREEN		**COSINE**		**COCITATIONS**		**PAGERANKOFLINKS**	
Pierre de Fermat	*Germany*	*Pierre de Fermat*	*Germany*	*Pierre de Fermat*	*Germany*	*Pierre de Fermat*	*Germany*
1. Pierre de Fermat 2. Mathematics 3. Probability theory 4. Fermat's Last Theorem 5. Number theory 6. Toulouse 7. Diophantine equation 8. Blaise Pascal 9. Fermat's little theorem 10. Calculus	1. Germany 2. Berlin 3. France 4. Austria 5. German language 6. Bavaria 7. World War II 8. German Democratic Republic 9. European Union 10. Hamburg	1. Pierre de Fermat 2. ENSICA 3. Fermat's theorem 4. International School of Toulouse 5. École Nationale Supérieure d'Électronique, d'Électrotechnique… 6. Languedoc 7. Hélène Pince 8. Community of Agglomeration of Greater Toulouse 9. Lilhac 10. Institut d'études politiques de Toulouse	1. Germany 2. History of Germany since 1945 3. History of Germany 4. Timeline of German history 5. States of Germany 6. Politics of Germany 7. List of Germany-related topics 8. Hildesheimer Rabbinical Seminary 9. Pleasure Victim 10. German Unity Day	1. Pierre de Fermat 2. Leonhard Euler 3. Mathematics 4. René Descartes 5. Mathematician 6. Gottfried Leibniz 7. Calculus 8. Isaac Newton 9. Blaise Pascal 10. Carl Friedrich Gauss	1. Germany 2. United States 3. France 4. United Kingdom 5. World War II 6. Italy 7. Netherlands 8. Japan 9. 2005 10. Category:Living people	1. France 2. 17th century 3. March 4 4. January 12 5. August 17 6. Calculus 7. Lawyer 8. 1660 9. Number theory 10. René Descartes	1. United States 2. United Kingdom 3. France 4. 2005 5. Germany 6. World War II 7. Canada 8. English language 9. Japan 10. Italy
Mark: 7.0/10	Mark: 5.5/10	Mark: 2.9/10	Mark: 7.4/10	Mark: 5.4/10	Mark: 2.1/10	Mark: 2.5/10	Mark: 1.1/10

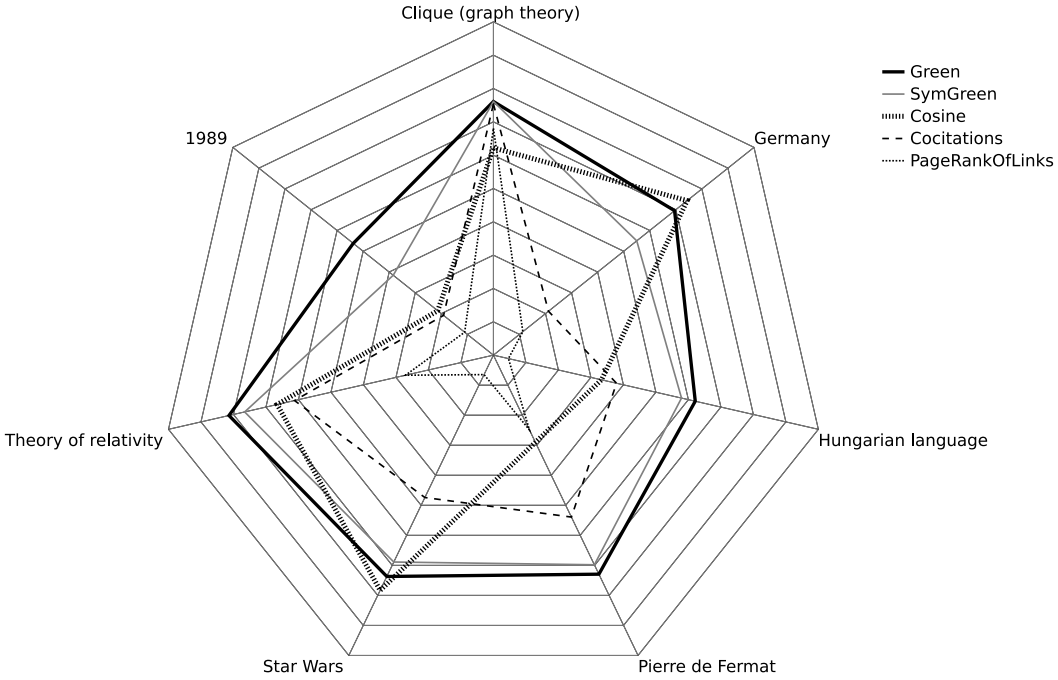

Figure 1: Radar chart of the average marks given to each method on the various base articles.

Table 3: Evaluation results. For each method, the following figures are given: average mark, averaged on all articles; 90% Student's t-distribution confidence interval; article-to-article standard deviation; evaluator-to-evaluator standard deviation; global count of $10/10$ marks; average mark for each article.

	GREEN	**SYMGREEN**	**COSINE**	**COCITATIONS**	**PAGERANKOFLINKS**
Average mark	7.0	6.3	5.2	4.5	2.2
90% confidence interval	$\pm$0.3	$\pm$0.3	$\pm$0.3	$\pm$0.3	$\pm$0.2
Article std. dev.	0.9	1.3	2.2	1.9	2.0
Evaluator std. dev.	1.7	1.7	1.9	2.0	1.6
Number of 10/10	25	10	12	9	4
Clique (graph theory)	7.6	7.6	6.2	7.5	6.8
Germany	7.0	5.5	7.4	2.1	1.1
Hungarian language	6.2	5.8	3.3	3.8	0.5
Pierre de Fermat	7.3	7.0	2.9	5.4	2.5
Star Wars	7.4	6.9	7.8	4.7	0.6
Theory of relativity	8.1	7.7	6.7	6.1	2.7
1989	5.4	3.8	2.1	1.9	1.1

PageRank values whatever the base article, except on articles with very few links.

Related Work

To our knowledge, this is the first use of discrete Green measures in the field of information retrieval on graphs or hyperlinked structures.

The relationship between Green measures and topic-sensitive PageRank (Haveliwala 2003) has been discussed above. Note that, in addition to the mathematical differences, the purpose is not the same: in the case of topic-sensitive PageRank, classical keyword Web search focused on a specific part of the Web, with a measure of topic-wise importance; in our case, a measure of similarity unmarred by global PageRank values, and a definition of *conceptual neighborhoods* in a graph.

The problem of finding related nodes on the World Wide Web is not new. In his original well-known paper about *hubs* and *authorities* (Kleinberg 1999), Kleinberg suggests using authorities in a focused subgraph in order to compute *similar-page queries*; apart from the use of authorities instead of PageRank, this is very similar to **LOCALPAGERANK**, which performs poorly on Wikipedia. In (Dean & Henzinger 1999), the authors present two different approaches for finding related pages on the Web: the *Companion* algorithm, which uses authorities scores in a local subgraph, and a cocitation-based algorithm.

In the specific case of Wikipedia, (Adafre & de Rijke 2005) uses a cocitation approach to identify missing links. We saw that **COCITATIONS** fared much worse than **GREEN** in our experiment. *Synarcher* (Krizhanovsky 2005) is a program for synonym extraction in Wikipedia, relying on authority scores in a local subgraph (comparable to **LOCALPAGERANK**) together with the information provided by Wikipedia's category structures. In (Grangier & Bengio 2005) a technique is presented to modify a classical text mining similarity measure (based on full textual content) by taking the hyperlinks into account using machine learning; no application to the problem of finding related pages is given.

Conclusion and perspectives

We showed how to use Green measures for the extraction of related nodes in a graph. This is a generic, parameter-free algorithm, which can be applied *as is* to any directed graph. We have described and implemented in a uniform way other classical approaches for finding related nodes. Finally, we have carried out a user study on the example of the graph of Wikipedia. The results show that the Green method has three advantages: 1. Its average performance is high, significantly above that of all other methods. 2. It is *robust*, never showing a bad performance on an article. 3. It is able to unveil semantic relationships not found by the other methods.

There is much room for extensions and improvements, either on the theoretical or the application side. For example it is easy to design variations on the Green method using standard variations on PageRank, such as HITS (Kleinberg 1999). Also, there is a continuous interpolation between **GREEN**, which follows only forward links, and **SYMGREEN**, which is bidirectional and tends to broaden the range of results (and is probably more robust). This could be used as a "specificity/generality" cursor.

A strong point of the methods presented here is that they rely only on the graph structure. It is very likely that, in the specific case of Wikipedia, we can improve performance by taking into account the textual content of the articles, the categories, some templates... although the raw method already performs quite well.

An obvious application is to try the Green method on the Web graph; this requires much more computational power, but seems feasible with large clusters of PCs. More generally, the method could be directly applied to any other context featuring associative networks.

Acknowledgments

We would like to thank all 67 participants who took the time to evaluate the different methods, and Serge Abiteboul for his feedback on the topic.

References

Adafre, S. F., and de Rijke, M. 2005. Discovering missing links in Wikipedia. In *Workshop on Link Discovery: Issues, Approaches and Applications*.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1–7):107–117.

Dean, J., and Henzinger, M. R. 1999. Finding related pages in the World Wide Web. *Computer Networks* 31(11–16):1467–1479.

Duffy, D. G. 2001. *Green's functions with applications*. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL.

Grangier, D., and Bengio, S. 2005. Inferring document similarity from hyperlinks. In *Conference on Information and Knowledge Management*.

Haveliwala, T. H. 2003. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering* 15(4):784–796.

Kemeny, J. G.; Snell, J. L.; and Knapp, A. W. 1966. *Denumerable Markov chains*. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London.

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

Krizhanovsky, A. A. 2005. Synonym search in Wikipedia: Synarcher. `arXiv:cs.IR/0606097`.

Norris, J. R. 1997. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

Ollivier, Y., and Senellart, P. 2007. Finding related pages using Green measures: An illustration with Wikipedia, companion website.
`http://pierre.senellart.com/wikipedia/`.

Salton, G., and McGill, M. 1984. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.