DirectoryRank: Ordering Pages in Web Directories

Vlassis Krikos
Computer Engineering Dept.
Patras University, Greece
krikos@ceid.upatras.gr

Sofia Stamou
Computer Engineering Dept.
Patras University, Greece
stamou@ceid.upatras.gr

Pavlos Kokosis
Computer Engineering Dept.
Patras University, Greece
kokosis@ceid.upatras.gr

Alexandros Ntoulas
Computer Science Department UCLA, USA
ntoulas@cs.ucla.edu

Dimitris Christodoulakis
Computer Engineering Department
Patras University, Greece
dxri@upatras.gr

ABSTRACT
Web Directories are repositories of Web pages organized in a hierarchy of topics and sub-topics. In this paper, we present DirectoryRank, a ranking framework that orders the pages within a given topic according to how informative they are about the topic. Our method works in three steps: first, it processes Web pages within a topic in order to extract structures that are called lexical chains, which are then used for measuring how informative a page is for a particular topic. Then, it measures the relative semantic similarity of the pages within a topic. Finally, the two metrics are combined for ranking all the pages within a topic before presenting them to the users.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: information filtering, retrieval models; H.3.m [**Information Systems**]: Miscellaneous

General Terms
Algorithms, Design, Experimentation, Measurement

Keywords
Web Directory, semantic similarity, ranking

1. INTRODUCTION
A Web Directory is a repository of Web pages that are organized in a topic hierarchy. Typically, Directory users locate the information sought simply by browsing through the topic hierarchy, identifying the relevant topics and finally examining the pages listed under the relevant topics. Given the current size and the high growth rate of the Web [10], a comprehensive Web Directory may contain thousands of pages within a particular category. In such a case, it might be impossible for a user to look through all the relevant pages within a particular topic in order to identify the ones that best represent the current topic. Practically, it would be more time-efficient for a user to view the Web pages in order of importance for a particular topic, rather than go through a large list of pages.

One way to alleviate this problem is to use a ranking function which will order the pages according to how "informative" they are of the topic that they belong to. Currently, the Open Directory Project [3] lists the pages within a category alphabetically, while the Google Directory [1] orders the pages within a category according to their PageRank [11] value on the Web. While these rankings can work well in some cases, they do not directly capture the closeness of the pages to the topic that they belong to.

In this paper, we present DirectoryRank, a new ranking framework that we have developed in order to alleviate the problem of ranking the pages within a topic based on how "informative" these pages are to the topic. DirectoryRank is based on the intuition that the quality (or informativeness) of a Web page with respect to a particular topic is determined by the amount of information that the page communicates about the given topic, relative to the other pages that are categorized in the same topic. Our method takes as input a collection of Web pages that we would like to rank along with a Web Directory's topic hierarchy that we would like to use. At a high level, our method proceeds as follows: first, we identify the most important words inside every page and we link them together, creating "lexical chains". We then use the topic hierarchy and the pages' lexical chains to compute the "relatedness" (or importance) of the pages to each of their corresponding topics. Having determined the pages' topic importance, we measure the relative semantic similarity among the pages that relate to the same topic. The semantic similarity indicates the amount of content that important pages in some topic share with each other. Finally, we employ our DirectoryRank algorithm that uses the topic importance scores in conjunction with the semantic similarities of the pages in order to compute the ranking order of the pages within a Directory topic.

In order to study the effectiveness of DirectoryRank in identifying the most informative pages within a particular topic, we applied our method to the ranking of 318,296 Web pages listed in 156 topics in the Google Directory. We have compared the rankings induced by DirectoryRank to the rankings induced by PageRank for the pages listed in those 156 topics. Our comparison reveals that the two rankings have different merits and thus they are useful in different tasks. To delve into the two rankings' effectiveness and investigate which is more useful for ordering pages in Directories' topics, we conducted a user study, where we asked a group of individuals to compare the rankings delivered by PageRank to the rankings delivered by DirectoryRank, and indicate which of the two is deemed as more useful. Our results show that, in most cases, the users perceived DirectoryRank to be more topic-informative than PageRank.

The rest of the paper is organized as follows: We start our discussion in Section 2 with a brief introduction to PageRank, which is currently employed by the Google Directory in order to rank pages. In Section 3, we briefly present the topic hierarchy that we use in our study as well as the process we follow for representing Web pages into lexical chains. We also show how we explore the topic

WIDM'05, November 5, 2005, Bremen, Germany

hierarchy and the pagers' lexical chains for measuring the pages' topic-importance and semantic similarities values. Finally, we present how our DirectoryRank metric employs the above values for measuring how informative Web pages are with respect to some topics and rank them accordingly. In Section 4, we experimentally study the effectiveness of DirectoryRank, by comparing its performance to PageRank. We revise related work in Section 5 and we conclude the paper in Section 6.

2. OVERVIEW OF PAGERANK

In this section, we briefly explain the main intuition of PageRank, a metric that was primarily invented for ranking pages within the Google Search Engine and that is currently used within the Google Directory for ordering Web pages. For a more elaborate overview on PageRank, we refer the reader to the work of [11]. The intuition of PageRank metric is that a page on the Web is important if there are a lot of other important pages pointing to it. That is, if a page *p* has many links from other important pages, we may conclude that this page is interesting to many people and that it should be considered as being "important" or of "good" quality. Similarly, if an important page has links to other pages, we expect that part of its quality is transferred to the pages it points to, which in turn become of increased significance/quality. Roughly, PageRank PR(*p*) defines the importance of page *p* to be the sum of the importance of the pages that endorse *p*. At a high level, PageRank is calculating the probability that a "random surfer" is looking at a given page at a given point of time. The "random surfer" is a mathematical model that emulates the behavior of a user that, given a page, follows an outgoing link from that page at random. Formally, given a page p_i that has incoming links from the pages p_1, …, p_n and let c_j be the number of out-links from p_j, the PageRank of p_i is given by:

$$PR(p_i) = d + (1 - d) \left[PR(p_1) / c_1 + ... + PR(p_n) / c_n \right]$$

where *d* corresponds to the probability that the random surfer will get bored and his next visit will be a completely random page, and 1-*d* corresponds to the probability that the page the random surfer will pick for his next visit is an outgoing link of the current page.

3. DIRECTORY RANK

The ranking of a Web page within a particular topic intuitively depends on two criteria: (i) the importance of the page for the underlying topic. This criterion helps us identify the most important pages out of the several ones that may lie within a particular topic. (ii) the semantic correlation of a page relative to other important pages in the same topic. This criterion helps us rank pages relative to each other within a topic. For measuring the importance of a Web page in some topic, we explore a subject hierarchy that we have built in the course of an earlier study [13] and we use the lexical chaining technique for identifying the most important words inside the page.

We start our discussion with a presentation of the topic hierarchy that we use in our work (Section 3.1) and we describe the process we follow for representing Web pages into lexical chains (Section 3.2). We also explain how we utilize the topic hierarchy and the pages' lexical chains for measuring the pages' importance to the hierarchy's topics (Section 3.2.1). The contribution of our work lies in the exploitation of the topic hierarchy and the lexical chains that we generate for representing Web pages in order to compute the semantic similarities between the pages that are important in some topics. Moreover, we have developed a novel framework, which employs the pages' topic importance and semantic similarity measures for ranking pages inside Directory topics.

3.1 The Topic Hierarchy

The main intuition in our DirectoryRank metric is that topic relevance estimation of a Web page relies on the page's lexical coherence, i.e. having a substantial portion of words associated with the same topic. To capture this property, we adopt the lexical chaining approach: for every Web page we generate a sequence of semantically related terms, known as lexical chain. In our approach of representing Web pages into lexical chains, we adopt the method reported in [6], which uses WordNet [5] as the knowledge base for providing information about the semantic relations that exist between the words in a text. A detailed description of the lexical chains' generation process is given in Section 3.2. Before that, we present the topic hierarchy that we use for determining the topics that are associated with the contents (i.e. words) of Web pages.

Since we are mainly interested in measuring the Web pages' importance in the context of Web Directories, we decided to demonstrate the usefulness of our DirectoryRank metric in ordering Web pages in the topics currently used in a real Web Directory. To that end, we applied DirectoryRank to the main topics used in the Google Directory. Google Directory provides a hierarchical listing of Web pages categorized by topic and reuses the data maintained by the Open Directory Project. Moreover, since DirectoryRank relies on the Web pages' lexical chains rather than their entire contents for measuring the pages' importance to particular topics and since lexical chain generation is dependent on WordNet, we decided to enrich the top level (main) topics of the Google Directory with their respective WordNet lexical hierarchies.

The first step we took for enriching the Google topics with WordNet data was to examine the compatibility between these topics and the topics used to annotate WordNet's concepts with domain information. Note that the topic information that exists in the labels of WordNet's contents is taken from the freely available Suggested Upper Merged Ontology (SUMO) [4] and the MultiWordNet Domains (MWND) [2]. Due to space limitations, here we present a summary of our approach into enriching the Google topics with WordNet hierarchies. A detailed description of the process we followed for appending to the Google top level topics their corresponding WordNet hierarchies is given in [13]. In brief, we located the Google's top level topics among the topics used in either SUMO or MWND for annotating WordNet concepts. Out of the 17 Google topics, 13 topics (shown in Table 1) are used for labeling WordNet concepts with topic information. To each of those 13 topics, we integrated their corresponding sub-topics that we acquired from either SUMO or MWND. The sub-topic integration was performed automatically, simply by following WordNet's hyper/hyponymy links. At the end of this process, we came down to a hierarchy of 489 sub-topics, which are organized into the 13 top level topics that we used from Google Directory.

Table 1. The Hierarchy's First Level Topics

First Level Topics	
Arts	News
Sports	Society
Games	Computers
Home	Reference
Shopping	Recreation
Business	Science
Health	

In Section 3.4, we will demonstrate how to use our topic hierarchy for automating the task of ranking pages within topical categories.

3.2 Measuring Web Pages' Topic Importance

The computational model that we adopted for generating lexical chains is presented in the work of Barzilay [6] and it generates lexical chains in a three step approach: (i) it selects a set of candidate terms[1] from a page, (ii) for each candidate term, it finds an appropriate chain relying on a relatedness criterion among members of the chains, and (iii) if such a chain is found, it inserts the term in the chain. The relatedness factor in the second step is determined by the type of WordNet links that connect the candidate term to the terms stored in existing lexical chains. Figure 1 illustrates an example of the lexical chain generated for a text containing the candidate terms: system, network, sensor, weapon, missile, surface and net. The subscript *si* denotes the id of the word's sense within WordNet.

Lexical chain
system $_{s6}$ → network $_{s4}$
system $_{s6}$ → sensor $_{s1}$
system $_{s6}$ → weapon $_{s2}$ → missile $_{s1}$
system $_{s6}$ → surface $_{s1}$ → net $_{s2}$

Figure 1. An example of a lexical chain.

Having generated lexical chains, we disambiguate the sense of the words inside every chain by employing the scoring function *f* introduced in [12], which indicates the probability that a word relation is a correct one.

Given two words, w_1 and w_2, their scoring function *f* via a relation *r*, depends on the words' association score, their depth in WordNet and their respective relation weight. The association score (*Assoc*) of the word pair (w_1, w_2) is determined by the words' co-occurrence frequency in a corpus that has been previously collected. In practice, the greater the association score between a word pair w_1 and w_2 is, the greater the likelihood that w_1 and w_2 refer to the same topic. Formally, the (*Assoc*) score of the word pair (w_1, w_2) is given by:

$$Assoc(w_1, w_2) = \frac{\log (p(w_1, w_2) + 1)}{N_s(w_1) \square N_s(w_2)}$$

where $p(w_1,w_2)$ is the corpus co-occurrence probability of the word pair (w_1,w_2) and $N_s(w)$ is a normalization factor, which indicates the number of WordNet senses that a word w has. Given a word pair (w_1, w_2) their *DepthScore* expresses the words' position in WordNet hierarchy and is defined as:

$$DepthScore(w_1, w_2) = Depth(w_1)^2 \square Depth(w_2)^2,$$

where *Depth* (w) is the depth of word w in WordNet. Semantic relation weights (*RelationWeight*) have been experimentally fixed to 1 for reiteration, 0.2 for synonymy and hyper/hyponymy, 0.3 for antonymy, 0.4 for mero/holonymy and 0.005 for siblings. The scoring function *f* of w_1 and w_2 is defined as:

$$f_s(w_1, w_2, r) = Assoc(w_1, w_2) \square DepthScore(w_1, w_2) \square \text{Re}lationWeight(r)$$

The value of the function *f* represents the probability that the relation type *r* is the correct one between words w_1 and w_2. In order to disambiguate the senses of the words within lexical chain C_i we calculate its score, by summing up the f_s scores of all the words w_{j1} w_{j2} (where w_{j1} and w_{j2} are successive words) within the chain C_i. Formally, the score of lexical chain C_i, is expressed as the sum of the score of each relation r_j in C_i.

$$Score(C_i) = \sum_{r_j \, in \, C_j} f_s(w_{j1}, w_{j2}, r_j)$$

[1] As candidate terms, we use nouns and proper names because they convey the vast majority of conceptual information in texts.

Eventually, in order to disambiguate we will pick the relations and senses that maximize the *Score* (C_i) for that particular chain. In estimating the importance of a Web page p_i in some Directory's topic T_k our first step is to identify which node within the hierarchy (see Section 3.1) corresponds to topic T_k of the page.

3.2.1 Topic-Importance Scoring

Once the topic of a page is located among the hierarchy's topics, we map the words in the page's lexical chain to the WordNet nodes under that particular topic. Recall that upon lexical chain generation, words are disambiguated ensuring that every word inside a page is mapped to a single word within the WordNet hierarchy. We then determine the importance of a page p_i to topic T_k by counting the number of words in the lexical chain of p_i that are subsumed by T_k in the hierarchy's graph. The topic importance of a page is given by a *Relatedness Score* (*RScore*), which indicates how relevant a page is for a given topic. Formally, the relatedness score of a page p_i (represented by the lexical chain C_i) to the hierarchy's topic T_k is defined as the product of the page's chain Score (C_i) and the fraction of words in the page's chain that are descendants of T_k. Formally, the *RScore* is given by:

$$RScore(C_i, T_k) = \frac{Score(C_i) \bullet \left| \text{common } C_i \text{ and } T_k \text{ elements} \right|}{\left| C_i \text{ elements} \right|}$$

The denominator is used to remove any effect the length of a lexical chain might have on *RScore* and ensures that the final score is normalized so that all values are between 0 and 1, with 0 corresponding to no relatedness at all and 1 indicating the page that is highly expressive of the page's topic. The *RScore* of a page to a specific topic captures the importance of the page in the given topic.

3.3 Semantic Similarity Scoring

The relatedness score metric that we have just presented can serve as a good indicator for identifying the most important pages within a topic. However, the *RScore* metric does not capture the amount of common content that is shared between the Web pages in a topic. This is important in the cases where our topic-importance scoring gives a low score for some pages but, at the same time, these pages are very similar to other pages with high topic-importance scores. In order to accommodate for this scenario, we now show how to compute the semantic similarities among the pages that are listed in the same Directory topic. Semantic similarity is indicative of the pages' semantic correlation and helps in determining the ordering of the pages that are deemed important in some topic. Our DirectoryRank metric employs the Web page's topic-importance scores and their semantic similarities to determine their ranking order inside some Directory topics and is presented in the next section.

In order to estimate the Web pages' semantic similarity, we compare the elements in a page's lexical chain to the elements in the lexical chains of the other pages in a Directory topic. We assume that if the chains of two Web pages have a large number of elements in common, then the pages are correlated to each other. To compute similarities between pages, p_i and p_j that are categorized in the same topic, we first need to identify the common elements between their lexical chains, represented as PC_i and PC_j, respectively. First, we use WordNet to augment the elements of the lexical chains PC_i and PC_j with their synonyms. Chain augmentation ensures that pages of comparable content are not regarded unrelated, if their lexical chains contain distinct, but semantically equivalent elements. The augmented elements of PC_i and PC_j are defined as:

$$AugElements(PC_i) = C_i \bigcup Synonyms(C_i)$$
$$AugElements(PC_j) = C_j \bigcup Synonyms(C_j)$$

where, *Synonyms* (C_i) denotes the set of the hierarchy's concepts that are synonyms to any of the elements in C_i and *Synonyms* (C_j) denotes the set of the hierarchy's concepts that are synonyms to any of the elements in C_j. The common elements between the augmented lexical chains PC_i and PC_j are determined as:

$$ComElements(PC_i, PC_j) = AugElements(PC_i) \bigcap AugElements(PC_j)$$

We formally define the problem of computing pages' semantic similarities as follows: if the lexical chains of pages p_i and p_j share elements in common, we produce the correlation look up table with tuples of the form <*AugElements* (PC_i), *AugElements* (PC_j), *ComElements*>. The similarity measurement between the lexical chains PC_i, PC_j of the pages P_i and P_j is given by:

$$\sigma_S(PC_i, PC_j) = \frac{2 \bullet \left| ComElements(PC_i, PC_j) \right|}{\left| AugElements(PC_i) \right| + \left| AugElements(PC_j) \right|}$$

where, the degree of semantic similarity is normalized so that all values are between zero and one, with 0 indicating that the two pages are totally different and 1 indicating that the two pages talk about the same thing.

3.4 DirectoryRank Scoring

Pages are sorted in Directory topics on the basis of a DirectoryRank metric, which defines the importance of the pages with respect to the particular topics in the Directory. DirectoryRank (*DR*) measures the quality of a page in some topic by the degree to which the page correlates to other informative/qualitative pages in the given topic. Intuitively, an informative page in a topic, is a page that has a high relatedness score to the Directory's topic and that is semantically close (similar) to many other pages in that topic. *DR* defines the quality of a page to be the sum of its topic relatedness score and its overall similarity to the fraction of pages with which it correlates in the given topic. This way, if a page is highly related to topic *D* and also correlates highly with many informative pages in *D*, its *DR* score will be high. Formally, consider that page p_i is indexed in Directory topic T_k with some *RScore* (p_i, T_k) and let $p_1, p_2, \ldots, p_n$ be pages in T_k with which p_i semantically correlates with scores of σ_s (PC_1, PC_i), σ_s (PC_2, PC_i),…, σ_s (PC_n, PC_i), respectively. Then, the DirectoryRank (*DR*) of p_i is given by:

$$DR(p_i, T_k) = RScore(p_i, T_k) + [\sigma_s(PC_1, PC_i) + \sigma_s(PC_2, PC_i) + \ldots\ldots + \sigma_s(PC_n, PC_i)] / n$$

where n corresponds to the total number of pages in topic T_k with which p_i semantically correlates.

4. EXPERIMENTAL SETUP

To measure the potential of our DirectoryRank metric in delivering topic-informative rankings, we conducted an experiment, where we studied the effectiveness of *DR* in prioritizing the most informative pages in some Directory's topics. To obtain perceptible evidence of DirectoryRank's efficiency in a practical setting, we applied our *DR* metric to a set of Web pages listed in a number of topics in Google Directory and we compared the rankings induced by DirectoryRank to the rankings that Google Directory delivers for the same set of pages and topics. In Section 4.1 we explain how we selected the pages for our study, while in Section 4.2 we present the similarity measure that we used for comparing the rankings induced by DirectoryRank to the rankings delivered by PageRank, and we give obtained results. Moreover, to delve into the behavior of DirectoryRank we carried out a user study, presented in Section 4.3.

4.1 Experimental Dataset

In selecting our experimental data, we picked pages that are categorized in those topics in Google Directory, which are also present in our hierarchy. Recall that Google Directory is a replica of the Dmoz Directory, from which we borrowed our hierarchy's 13 top-level topics. Out of all the sub-topics organized in those 13 top-level topics in Google Directory, 156 were represented in our hierarchy. Having determined the topics, whose set of ranked pages would be compared, we downloaded a total number of 318,296 pages, categorized in one of the 156 selected topics, which in turn are organized into the 13 top-level topics. Table 2 shows the distribution of the experimental pages in the top level topics in Google Directory.

Table 2. Statistics on the experimental data

Category	# of documents	# of sub-topics
Arts	28,342	18
Sports	20,662	26
Games	11,062	6
Home	6,262	7
Shopping	52,342	15
Business	60,982	7
Health	23,222	7
News	9,462	4
Society	28,662	14
Computers	35,382	13
Reference	13,712	10
Recreation	8,182	20
Science	20,022	9
Total	318,296	156

Since we were interested in comparing DirectoryRank with PageRank, in the context of ranking Web pages in Directory topics, we recorded for the downloaded Web pages their relative ranking order in Google Directory in each of the 156 selected topics. We then stored the downloaded pages in a secondary index, maintaining their relative PageRank rankings. To compute the *DR* values for every experimental page, we initially processed the downloaded pages in order to generate and score their lexical chains. For every page, we first computed its *RScore* to the topic in which it is assigned in Google Directory, and then we computed the semantic similarity (σ_s) for every pair of pages listed in each topic. Lastly, using the above two scores (i.e. semantic similarity and topic relatedness), we computed for every Web page its DirectoryRank (*DR*) value and we sorted the pages listed within each of the topics, so that pages with higher *DR* scores in some topic are prioritized among the set of topic related pages. Using the above data, we evaluated the effectiveness of our DirectoryRank metric in ordering Web pages inside the Directory's topics.

4.2 Overlap of DirectoryRank and PageRank

To investigate whether there is any similarity between the rankings induced by DirectoryRank and the rankings delivered by PageRank for our experimental pages in the 156 topics in Google Directory, we used the *OSim* measure, reported in the work of [9], which indicates the degree of overlap between the top *n* URLs of the two rankings. Formally, the overlap of two ranked lists A and B (each of size *n*) is given by: $OSim(DR, PR) = |A \cap B| / n$

Using the above formula, we computed for each of the 156 topics the overlap between the pages ranked in the top n=10 positions for that topic by *DR* and *PR* respectively. Afterwards, we first com-

puted the average similarity between the two induced rankings for each of the 156 selected topics, and then the average similarity between the two induced rankings for each of the 13 top-level topics. To compute the average similarity between *DR* and *PR* for a top level topic *T*, we summed the average similarity of all sub-topics in *T* and we divided by the number of sub-topics that *T* has. Table 3 gives the average similarity scores between *DR* and *PR* for each of the 13 top-level topics examined in our experiment.

Table 3. Average similarity of rankings for the top level topics

Category	*OSim*
Arts	0.038
Sports	0.019
Games	0.030
Home	0.057
Shopping	0.013
Business	0.028
Health	0.057
News	0.100
Society	0.043
Computers	0.046
Reference	0.020
Recreation	0.025
Science	0.044

Obtained results demonstrate that there is little average overlap between the top 10 results for the two rankings. Note that for some topics we compared the overlap between *DR* and *PR* for a larger set of pages (e.g. *n*=20 and *n*=30) and we found that the *OSim* score of the two rankings increases, albeit slightly, as the size of *n* grows. For example in the topic Sports, the *OSim* between *DR* and *PR* for *n*=10 is 0.019, whereas for *n*=20 the *OSim* score is 0.023 and for *n*=30, *OSim* is 0.028. Our results show that even the pairs with the greatest similarity among all pairs examined (e.g. the rankings delivered for the News topic), according to the *OSim* measure, have little in common. Despite the usefulness of the *OSim* measure for making rough estimations about the ability of the two ranking schemes in identifying the same top pages with respect to some topics, it cannot directly capture which ranking is more useful for ordering pages in Directory topics. This is because *OSim* does not indicate the degree to which the relative orderings of the top *n* pages of two rankings are in agreement. Having established that PageRank and DirectoryRank order Web pages substantially differently, we proceed to investigate which of these rankings is better for ordering Web pages in Directory topics. To that end, we carried out a user study, reported next.

4.3 DirectoryRank Performance

To determine which of the two ranking measures, namely *DR* and *PR*, is perceived as more useful by Web users for organizing pages in Web Directories, we carried out a user study. From our sample data, we picked the top 10 pages listed in 7 randomly selected topics (out of the 156 topics examined) and we recruited 15 postgraduate volunteers from our school. Table 4 lists the 7 topics selected.

For each topic, the volunteer was shown 2 result rankings; one consisted of the top 10 pages for the topic ranked with *DR*, and the other consisted of the top 10 pages for the topic ranked with *PR*. For each topic, the volunteer was asked to read the pages in both lists and indicate which of the two rankings, in their opinion, is more "useful" overall for communicating information about the topic. Volunteers were not told anything about how either of the rankings was generated. In order to avoid misinterpretations while analyzing the user's selection preferences, we asked from the users to indicate their descriptive selections directly. More specifically, we presented to our participants the following choices and we asked them to indicate for which of the following reasons they selected one ranking over the other for each of the topics examined.

Table 4. Experimental Topics

Experimental Topics	
$\mathbf{T_1}$	Crime
$\mathbf{T_2}$	Photography
$\mathbf{T_3}$	Water Sports
$\mathbf{T_4}$	Radiology
$\mathbf{T_5}$	Mechanics
$\mathbf{T_6}$	Econometrics
$\mathbf{T_7}$	Collecting

Reason R1. "I prefer this ranking because I obtained significant information about the topic from most of the pages". In our analysis, we interpret the ranking preferences established on this reason as "topic-informative" rankings.

Reason R2: "I prefer this ranking because I have seen most of the pages before and I liked them". We interpret the ranking preferences established on this reason as "popular" rankings.

We then compared the participants' descriptive selections for every topic with the final *DR*/ *PR* choices. This way we ensure that users' preferences would be accurately evaluated even if two volunteers had exactly the same descriptive selection, but they ended up casting that selection into different *DR*, *PR* rankings. As a final note, we also asked our volunteers to indicate their familiarity with the experimental topics, by characterizing as "familiar" or "unfamiliar" each of the topics examined. In our evaluation, we considered that one ranking was better than the other if at least 50% of the users selected it as more "useful". Table 5 shows the rankings selected by our subjects as more useful for each of the 7 examined topics. Every row corresponds to a separate user. The columns marked as T_i show what the preference of the user was for the particular topic. Under the T_i columns the keyword *DR* means that the user considered DirectoryRank as more useful for that topic, while PR means that the user deemed PageRank as more useful. The column marked as R on the right of a T_i column indicates the reason for which the user voted over the specified ranking. Table 6 summarizes the rankings preferred by the majority of the users for each of the topics.

Table 5. Rankings selected as more useful for each topic

User	T_1	R	T_2	R	T_3	R	T_4	R	T_5	R	T_6	R	T_7	R
#1	**DR**	1	**DR**	1	**DR**	1	**DR**	1	PR	2	**DR**	1	PR	2
#2	PR	2	**DR**	2	PR	2	**DR**	1	**DR**	1	**DR**	1	PR	2
#3	**DR**	1	**DR**	1	**DR**	1	**DR**	1	**DR**	2	**DR**	1	PR	2
#4	PR	1	PR	1	PR	2	**DR**	2	PR	2	PR	2	PR	1
#5	**DR**	1	PR	1	PR	2	PR	2	PR	2	**DR**	2	**DR**	1
#6	PR	2	**DR**	1	PR	2	**DR**	1	**DR**	1	**DR**	2	**DR**	1
#7	**DR**	2	PR	2	PR	1	**DR**	1	PR	2	**DR**	1	**DR**	1
#8	**DR**	1	**DR**	2	**DR**	1	**DR**	1	PR	1	**DR**	1	PR	2
#9	PR	2	**DR**	1	PR	2	PR	2	PR	2	**DR**	1	**DR**	2
#10	**DR**	1	**DR**	1	**DR**	1	**DR**	1	**DR**	1	**DR**	2	**DR**	2
#11	**DR**	1	**DR**	1	**DR**	1	**DR**	2	PR	2	PR	2	PR	2
#12	**DR**	1	**DR**	1	**DR**	1	PR	1	PR	2	**DR**	1	**DR**	1
#13	**DR**	2	PR	2	PR	1	**DR**	1	PR	2	**DR**	1	**DR**	1
#14	PR	2	**DR**	1	PR	2	**DR**	1	**DR**	1	**DR**	1	PR	2
#15	**DR**	1	**DR**	2	**DR**	1	**DR**	1	PR	1	**DR**	1	**DR**	1

Our survey results demonstrate that the majority of the users perceived in overall DirectoryRank as more useful in comparison to PageRank for ordering Web pages in the Directory's topics. This is attested by the fact that for most of the topics examined (for 5 out of the 7 topics), the majority of our subjects preferred *DR* over *PR*. A closer look at the obtained results indicates that the reason on which

our participants' based most of their *DR* selections, is Reason 1, which implies that the rankings delivered by *DR* are perceived as more topic-informative. Conversely, most of the users who liked better the rankings induced by *PR*, established their selection on Reason 2. This suggests that the usefulness of *PR* is not implied mainly by how informative a page is about a topic, but rather that it is substantially influenced by the page's popularity.

Table 6. Rankings preferred by the majority of users

Topic	Preferred by majority
$\mathbf{T_1}$ Crime	DirectoryRank
$\mathbf{T_2}$ Photography	DirectoryRank
$\mathbf{T_3}$ Water Sports	PageRank
$\mathbf{T_4}$ Radiology	DirectoryRank
$\mathbf{T_5}$ Mechanics	PageRank
$\mathbf{T_6}$ Econometrics	DirectoryRank
$\mathbf{T_7}$ Collecting	DirectoryRank

Moreover, although not reported here due to space limit, our survey results show that our participants' answers were not generally influenced by their familiarity or not with the underlying topics. This implies that our survey does not entail "topic-bias", since both rankings compared are applied to pages listed in the same topic.

5. RELATED WORK

There have been a number of studies trying to identify the best ranking order of the Web pages that are deemed to be relevant to a given query/topic. The most successful of these studies [8, 11] suggest the exploitation of the pages' links connectivity on the Web graph for measuring the pages' importance and rank them accordingly. The most widely known ranking metric that explores the pages' links structure for measuring their importance on the Web is PageRank. Currently, PageRank and its variations are used by most major Web Search Engines to rank the results that they return to Web users in response to their search requests. Despite PageRank's usefulness for ordering pages in the context of Search Engines, it is designed to measure the global importance of the pages on the Web, independent of any particular topics. However, the overall importance of the pages may be not a sufficient measure for ordering the pages inside Directories' topics, essentially because pages that are important in some topics may not be important in others, regardless of the number and structure of the links that may appear in those pages. To alleviate some of the inherent limitations of PageRank, a number of researchers designed new ranking metrics, which mainly rely on modifications of PageRank and are tailored for specific tasks. For example, [9] studies personalization of the PageRank metric by giving different weights to pages, [14] examine the local and the inter-site link structure in order to compute a global PageRank for Web pages, [7] introduce Hilltop, an algorithm which generates query-specific authority scores for improving rankings for popular queries. While most of these works mainly focus on improving the rankings delivered to Web users by measuring the Web pages' overall importance, in this paper we are more concerned about the topic importance of Web pages by measuring the pages' informativeness with respect to particular topics. In this scope, we perceive our work to be complementary to previous studies on personalized rankings [9]. Moreover, there exists prior work that explores the lexical chaining technique as a means for representing documents' contents [6, 12]. Recently, we employed the lexical chaining technique for the automatic classification of Web documents in topic hierarchies [13]. Our findings indicated the potential of lexical chains in successfully capturing the thematic content of Web pages. This motivated our work to use the lexical chains generated for a number of Web pages as a means for ordering pages within Directory topics. In the future we plan to investigate how our approach could benefit from other linguistic approaches, besides lexical chains.

6. CONCLUDING REMARKS

In this paper, we introduced DirectoryRank, a practical metric for determining how informative Web pages are for particular topics and ranking them accordingly. To evaluate the potential of DirectoryRank in ordering Web pages inside Directory topics, we conducted an experiment where we applied our DirectoryRank metric to order a set of pages listed within 156 topics in Google Directory and we compared the rankings induced by DirectoryRank to the rankings that PageRank delivers in Google Directory for the same set of pages and topics. In our study, we relied on the judgments made by 15 users to determine which ranking is perceived as more useful for Web Directories' users. Obtained results indicate that in overall users preferred DirectoryRank over PageRank for ordering Web pages inside the Directory's topics. Although it would probably require additional studies in order to evaluate the applicability of our method to Web Directories other than Google and assess DirectoryRank's usefulness to a larger user and categories base, we believe that our work can serve as the first step towards a topic-informative ranking metric within directories.

7. REFERENCES

[1] Google Directory http://dir.google.com/.

[2] MultiWordNet Domains http://wndomains.itc.it/.

[3] Open Directory Project http://dmoz.com/.

[4] Sumo Ontology http://ontology.teknowledge.com/.

[5] WordNet 2.0 http://www.cogsci.princeton.edu/~wn/.

[6] Barzilay R Lexical chains for text summarization. Master's Thesis, Ben-Gurion University, 1997.

[7] Bharat K and Mihaila G. Hilltop: a search engine based on expert documents: http://www.cs.toronto.edu/~georgem/ hilltop/.

[8] Kleinberg J. Authoritative sources in a hyperlinked environment. In Journal of the ACM, 46(5), 1999, 604-632.

[9] Haveliwala T. Topic sensitive PageRank. In Proceedings of the 11th WWW Conference, 2002, 517-526.

[10] Ntoulas A., Cho J. and Olston Ch. What's new on the web? The evolution of the web from a search engine perspective. In Proceedings of the 13th WWW Conference, 2004, 1-12.

[11] Page L., Brin S., Motwani R. and Winograd T. The PageRank citation ranking: Bringing order to the web. Available at http://dbpubs.stanford.edu:8090/pub/1999-66.

[12] Song Y.I., Han K.S. and Rim H.C. A term weighting method based on lexical chain for automatic summarization. In Proceedings of the 5th CICLing Conference, 2004, 636-639.

[13] Stamou S., Krikos V., Kokosis P., Ntoulas A. and Christodoulakis D. Web directory construction using lexical chains. In Proceedings of the 10th NLDB Conference 2005, 138-149.

[14] Wang Y. and DeWitt D. Computing PageRank in a distributed internet search system. In Proc. of the 30th VLDB Conf., 2004.