Hyperlink of men

Ricardo Kawase, Patrick Siehndel, Eelco Herder and Wolfgang Nejdl
Leibniz University of Hanover & L3S Research Center
Appelstrasse 9, 30167 Hannover, Germany
{*kawase, siehndel, herder, nejdl*}*@L3S.de*

***Abstract*—Hand-made hyperlinks are increasingly outnumbered by automatically generated links, which are usually based on text similarity or some sort of recommendation algorithm. In this paper we explore the current linking and appreciation of automatically generated links. To what extent do they prevail on the Web, in what forms do they appear, and do users think that generated link are just as good as human-created links? To answer these questions we first propose a model for extracting contextual information of a hyperlink. Second, we developed a hyperlink ranker to assigned relevance to each existing human generated link. With the outcomes of the hyperlink ranker, together with another two recommendation strategies, we performed a user study with over 100 participants. Results indicate that automated links are 'good enough', and even preferred in some user contexts. Still, they do not provide the deeper knowledge as expressed by human authors.**

I. Introduction

Wikipedia [1] defines the World-Wide Web as 'a system of interlinked hypertext documents accessed via the Internet'. This straightforward definition emphasizes the central role of hyperlinks on the Web, which allow and encourage authors to divide pieces of information in units that can be related in multiple manners. Hyperlinks - mostly simply called links - appear in many different forms on the Web. Links that are embedded in menus or navigation bars guide readers through a site or information domain. Within-text links are often associative links and point to related topics, elaboration or background information. With the advent of content management systems, menus and navigation bars are often generated automatically. By contrast, associative links were usually authored by hand. That is to say, until recently.

Traditional hand-made links got company of - or are replaced by - automatically generated links, based on text similarity or some sort of recommendation algorithm. These links may even be personalized - adapted to a user profile that the system may have. In many cases, there is no apparent visual difference between hand-made links and automatically generated links. Automatic generation of links saves Web authors of creating links themselves, and some voices claim that automated methods result in a more coherent and complete link structure.

[1]http://en.wikipedia.org/wiki/World_Wide_Web, retrieved 6 February 2012

At the same time, opponents raise their voices against automated links as well. Automated links cannot compete with the creative, serendipitous and associative ways in which human authors would think. Therefore, automated links may lead to some form of tunnel vision. To make things worse, personalized recommendations may lead users to only find what (the system thinks) they want to find - a phenomenon that has been pointed out by Eli Pariser in his TED talk 'Beware online filter bubbles' [2].

In this paper we explore the current practices and appreciation of automatically generated links. To what extent do they prevail on the Web and in what forms do they appear? And in particular: do users think that generated link are just as good as hand-made links - or perhaps even better? To answer these questions we set up an experiment in which we presented users Wikipedia articles with three different sets of links - one set based on manually created links, the two remaining sets were automatically created. On top of the hand-made links we developed a link ranker to extract and rank the relevant existing links of an article. The contribution of this paper is threefold:

- A model to exploit contextual features in hyperlink assigments.
- A link ranker that extracts and sort the most important hyperlinks of a document.
- The comparison between human-generated hyperlinks against automatically created ones.

After a discussion of background and related work in Section II, we explore the prevalence of automated and hand-made links in news sites in Section III. The insights from this exploration serve as a motivation for our user study on the linkage of Wikipedia articles. In Section IV we describe our approach for extracting hand-made and generating automated links to similar articles and authorative articles. The setup of the user study is presented in Section V, followed by the results of the study in Section VI. In Section VII we discuss our findings and design implications. The paper ends with some conclusions and final remarks.

II. Background and Related Work

The history of hypermedia dates back to Vanevar Bush [4], who envisioned a machine, the *memex*, which would allow

[2]http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles.html

users to relate information sources and their own insights in an associative manner, as an alternative for the librarian approach of indexing by alphabet, author or publication date. In the 1960s Douglas Engelbart developed his oN-Line System (NLS), which consisted of a hierarchically structured set of statements; each statement could cross-reference any other statement. In the same period Ted Nelson coined the term *hypertext* for these kinds of systems and started the Xanadu project [9].

Pre-Web and early-Web hypermedia systems mainly contained *associative links*. Associative links are still relatively common in Web sites, but in addition *structural navigation* is provided imposing one or more *hierarchies* on the document, grouped in menus and other kinds of navigation bars. The context information provided by these structural links is important for effective navigation, as each navigation process is inextricably tied to the structure of the document.

Generated links are a popular feature in current content management systems and professional Web sites. Many sites provide navigation bars with links to pages that have been just added, that are most read ('Most Popular') or most commented upon. More importantly, recommendation techniques are increasingly used for recommending articles similar to the current one [3]. The similarity may be calculated based on the contents of the articles, on collaborative filtering or a combination of these two. A third category of generated links are semantic links, which typically refer to glossaries or background articles related to the linked term. The latter category of links is extensively used on news sites and the technique receives mixed responses [3]: *"No thinking human would ever add these links; obviously, a human has programmed a computer to automatically insert them"*.

The online encyclopedia Wikipedia is one of the world's largest sources of information. Wikipedia has been subject of a vast body of research and several projects focused on augmenting its link structure by automatically generating links. Adafre and Rijke [1] introduced a method based on to identify similar pages and links that might be missing on a given page. They also motivated why link augmentation is still useful in the already rigorously linked ecosystem of Wikipedia. As an illustrative example they showed that out of 65 randomly chosen articles on singers, only 34 articles are linked to the concept 'singer'. Making use of statistics on co-citation and page title information, their approach was meant to reduce the heterogeneity introduced by the large number of content contributors.

An alternative approach to augment the link structure in Wikipedia is the Wikify system by Mihalcea and Csomai [6]. Part of the system concentrates on the detection of phrases from which links could or should be made, by considering the number of Wikipedia articles that already use the phrase as an anchor - normalized by the total number of articles that contain the phrase, with or without link. The algorithm excludes all n-grams that do not reach a certain threshold. The system is reported to reach a precision of 53% and a recall of 56%.

A more refined approach to link detection is reported by [8]. Where Wikify used a threshold for deciding whether a link would be relevant, Milne and Witten made use of a number of other features, including link probability, location of the link, relatedness of the articles, confidence of disambiguation and the generality of the linked articles. Several of these features were also used for the link disambiguation.

From the above discussion it becomes apparent that the traditional hand-crafted associative links are increasingly complemented - or even replaced - by automatically generated links. Apart from the structural links that guide users to the site content and lists of the most popular or latest pages, generated links are often used to connect pages with similar content, to point the reader to useful reference pages or to provide personalized recommendations for pages that a user is thought to be interested in.

A logical question that arises is whether and how automatically generated links serve as an alternative for human-authored links. Traditionally, hand-made links are considered useful, because they cover the associative way 'we may think'. By contrast, automated links are thought to link entities more consistently and reliably. At the end, it will not be a machine evaluation that decides upon this, but the world-wide population of Web users. Therefore, in this paper, we evaluate which type of hyperlinks users consider most useful: hand-made links or automatically generated links. In the next section, we explain the algorithms we used for generating links - inspired by and making use of the features discussed by [6] and [8], as briefly explained in this section.

III. Links on News Sites

To motivate our work, our first aim was to find out which types of links are commonly used on the Web. As a representative domain we have chosen news sites, for two reasons. First, news sites are popular destinations on the Web and due to the heavy competition between news sites, the design of these sites typically keeps up with current practices. Second, despite differences in design and target users, they constitute a relatively homogeneous genre, which allows for easier comparison.

On November 1st 2011, we accessed 27 news sites that frequently appear as news sources for Google News. On each site, we performed three different queries for news articles on the portal site - these queries were 'Obama', 'Greece' and 'Hannover'. From the search results we accessed the first five listed articles [4] and analyzed the types of links that

[3] http://www.slate.com/articles/news_and_politics/press_box/2008/04/links_that_stink.html

[4] On three sites access to the full content was restricted

Table I
THE HYPERLINKS CHARACTERISTIC IN EACH NEWS PORTAL. THE COLUMN BOXES SHOWS THE EXISTING RECOMMENDATIONS BOXES ON EACH ARTICLE PAGE AND THE INLINE COLUMN SHOWS THE HYPERLINKS THAT ARE INLINE THE ARTICLES' TEXT.

Source	Boxes	Inline
The New York Times	R,MV,MC	E
The Washington Post	R,MV	E,A
Houston Chronicle	L	S
Bloomberg L.P.	MV,L	E
Los Angeles Times	R,MV,MC,L	E
Reuters	R,MV,MC	E
Forbes	MV,L	A
Monsters and Critics.com	R,MV,L	-
guardian.co.uk	R,MV,L	E
Voice of America	R,MV,MC,L	-
International Herald Tribune	R,MV,MC	E
Boston Globe	na	na
Chicago Tribune	L	E
BBC News	R,MV	-
San Francisco Chronicle	R,MV,MC,L	E
CBS News	R,MV,MC,L	-
Times Online	na	na
Xinhua	R,MV,MC,L	-
Wall Street Journal	na	na
USA Today	R,MV,F,O	A
Fox News	R,MV,F,O	E
CNN	R,MV,O	A
Seattle Post Intelligencer	MV,MC,O	S
MSNBC	L	-
ABC News	MV	-
Daily Mail	L,F	-
The Times of India	R,MV,MC,L	S

R	Related	16	A	Article	4
MV	Most Viewed	20	E	Entity	10
MC	Most Commented	10	S	Search	3
L	Latest	14	na	Not Available	3
F	Featured	3			
O	Other	4			

were present in navigation boxes or bars and in the running text. These different types were categorized and labeled.

The results of this exercise are summarized in Table I. We counted a link category as a feature of the news site if it occurred in at least one of the fifteen articles that we inspected on each site.

In the left part of the table the link categories found in navigation boxes or bars are presented and their occurrence on the news sites that we examined.

- **R** - *Related*: A list of related articles, either manually created or automatically generated. The list may appear at the end of an article or in a separate box.
- **MV** - *Most Viewed*: The most popular or most read articles, not (necessarily) related to the current article.
- **MC** - *Most Commented*: A selection of articles that received most comments or that were often shared in social media or via email.
- **L** - *Latest*: A listing of articles that were most recently added to the site.
- **F** - *Featured*: A selection of featured articles, not (necessarily) related to the current article - presumably hand-picked by an editor.
- **O** - *Other*: A list of articles without a clear label or purpose, in many cases more or less related to the current article.

Most news sites display at least one navigation box - apart from the main menu structure. When only one box is present, this navigation box usually contains automatically generated links to the latest or most viewed articles. 'Most viewed' appears to be more commonly used than a list of latest articles, and it comes together with a list of most commented articles in half of the cases. Links to related articles seldom come as the only navigation box, which suggests that such a list is considered of secondary importance. In most cases it was not clear whether the related articles were manually picked or automatically generated, but the low prevalence of manually picked featured articles suggests a tendency toward automated links.

In the right part of the table you find the types of links found in the running text of the articles.

- **A** - *Article*: Within-text link that leads to a related article; the relation is typically expressed by the link text or a pop-up window.
- **E** - *Entity*: A link to a page that describes the entity mentioned in the link text. Such a link may be automatically generated or manually added.
- **S** - *Search*: A link that leads to a list of search results for the entity mentioned in the link text.

Seven news sites did not have any links in the running text of their articles. On the sites where within-text links were used, these links mainly pointed to background information on entities or to search results relating to the entity mentioned in the text. Only four news sites provided links to related articles - and this seemed to be highly dependent on the writing style of the author of the article and the time they spent on the articles.

Even though these numbers are based on a limited snapshot, they clearly suggest that manually generated links are not as commonly used as we might think they would. Instead, users are mainly guided by links that are generated by algorithms that serve as our information gatekeepers.

A natural question that follows these observations, is: are hyperlinks out of the box good enough? It is a likely assumption that hand-made links are more creative, serendipitous and associative, and therefore more valuable than links that are generated with fairly straightforward methods. But is this assumption correct? This question was our main motivation for setting up the user study that is described in the following sections.

IV. LINK GENERATION AND EXTRACTION

As we have seen in the related work and the analysis of links in news articles, the most common types of generated links that are article-specific are *recommendations* for similar articles and links to the *main entities* that are detected in the running text. These are the two types of generated links that we will evaluate in our experiment. In order to

compare these generated links with a set of human-created links, we developed a technique for *extracting* and ranking (man-made) links from running text.

A. L-Recommender

The first method, *L-Recommender*, is a link recommender that is based on textual similarity between articles. Similarity-based recommendations constitute a well-researched subdomain of content-based recommender systems. In our study, we used MoreLikeThis, a standard function provided by the Lucene search engine library [5].

MoreLikeThis calculates the relatedness of two documents by computing the number of overlapping words and giving them different weights based on TF-IDF [11]. MoreLikethis runs over the fields we specified as relevant for the comparison - in our case the title and the corpus of the articles -, and generates a term vector for each analyzed item (excluding stop-words). For the calculation of similar documents, the method only considered words that are longer than 2 characters and that appear at least 5 times in the source document. Also words that occur in less than 5 different documents are not taken into account for the calculation. For calculating the relevant documents, the method used the 15 most representative words (based on their TD-IDF values) and generates a query with these words. The ranking of the resulting documents is based on Lucene's scoring function which is based on the Boolean model of Information Retrieval and the Vector Space Model of Information Retrieval [12].

With this approach, terms that occur in fewer documents are considered to be more representative for the article and hence more relevant for finding related articles.

B. L-Detector

The second strategy is a method that involves 'detecting' phrases from which links could be made, similar to the approach followed by Mihalcea and Csomai[6] (see Section II). The *L-Detector* employs an authority-based approach; in our setup, we consider authorities to be the Wikipedia articles that have the largest number of incoming links. The L-Detector scans the whole content of the articles and outputs the top-n links to articles that have the largest number of incoming links.

In normal cases, these detected links would be visualized as within-text links to the corresponding entities - as is the case with entity links in news sites, as described in Section III. This would imply that the number of links needs to be balanced in such a way that the running text is not flooded with links. Since our evaluation setup will only take into account the top links of each proposed strategy, our main concern for this strategy is in terms of maximazing the precision: we only need the top-n most authorative detected links that are relevant to the article.

[5]http://lucene.apache.org/java/3_0_0/api/contrib-queries/org/apache/lucene/search/similar/MoreLikeThis.html

C. L-Extractor

Our final strategy addresses the human-created links that exist in the running text created by Wikipedia authors. For the purposes of our evaluation, the challenge of the *L-Extractor* is not only to extract the links, but also to *rank* them according to their (estimated) importance.

Given that we have the plain HTML of each article, the first step of extracting links is a straightforward parsing task. For assigning scores to each extracted hyperlink, we used a list of features that can be extracted from the contents and context of the links. For this purpose, we built a model in which page contents, hyperlinks and contextual information are related with one another. Traditionally, in hypertext systems, a link establishes a one-way connection between a page and its target. However, in our model we observe link assignment as a relation betweeen resources and contexts. Similar to the definition of a Folksonomy - which relates resources, users and tags -, we coin the term *Hypersonomy* for our model.

A Hypersonomy is a tuple $\mathbb{H} := (R, L, T, Y, C, Z)$, where R, L, T, C are finite sets of instances of resource (webpages), links, target resources and context information, respectively. Y defines the link assignment, which is a relation between R, L and T (i.e., $Y \subseteq R \times L \times T$). The context C is an open-concept that can be derived from observing features regarding the instances of the link assignment. Later in this subsection we describe six features that we exploit as context. Finally, Z defines the context assignment, which is a relation between Y and C (i.e., $Z \subseteq Y \times C$).

In a nutshell, the Hypersonomy simply defines the well-known hypertext structure, and additionally considers the context attached to the relation between the text and the hyperlinks.

To craft a *Hyperlink Ranker*, we built upon this concept, making use of the contextual information (features) embedded in the hyperlinks to assign scores to each of them. The Hyperlink Ranker is a function that takes as input the links assginments Y ($R \times L \times T$) of a page R_i together with its contextual information of the hyperlinks assignments Z, and assigns for each link $l_i \in L$ a value $v_i \in [0, 1]$ that is proportional to the importance of the link regarding the contextual features and the features of the target pages T.

We made use of the following contextual features to rank the links - many of them already discussed in the literature [6], [7], [8]:

(Link) Term Frequency. The first contextual feature we exploit is the TF score of the terms in the hyperlinks. As previously discussed in Section IV-A, the term frequency identifies the terms that better represent the current article. Following the same principle, the link term frequency is

assumed to be a measure for the importance of a link. The TF-scores are normalized between 0 and 1, as well as all the next strategies.

Location [7]. The location feature takes into account the observation made by David et. al [5] that terms that appear in the introduction of a document tend to be more important and relevant; the same applies for terms that occur in the conclusions of an article. Given the nature of on-screen reading [6], the latter observation is not valid in our context. Moreover, most Wikipedia articles are structured in a way that hyperlinks at the end of an article mainly point to external resources. Therefore, we take only the distance from the start of the article into account.

Additionally, we exploit features that are not contextualized in the hyperlink assignment, but are derived from the target article.

Length. The length of the target article is a feature that assigns higher scores to articles that have larger content. A well-elaborated Wikipedia article is usually the result of involvement of many collaborating editors and consequentiality a measure of relevance for the overall community. In previous work, Blumenstock demonstrated that counting just the number of words of an article is already a good feature for measuring article quality [2].

Generality. As shown by Milne et al. [7], it is more useful to present to the readers links that they are not likely to be familiar with, rather than providing links to very general, well-known pieces of information. As Wikipedia articles are structured in a hierarchical manner, we take as a measure for generality the number of sub-categories that one category has. Our link ranker assigns higher scores to links to articles that belong to rather specific categories. As the categories hierarchy follows a power distribution (few categories contain most of subcategories while most of them have few or none subcategories) we apply a logarithmic smoothing factor. The final score is given by the average of the scores of each category of the article.

Relatedness [7]. In the same manner as we calculated recommendations in Section IV-A, we use the similarity between the current article and the linked articles to increase the links' scores. One would expect that topics that are closer related to the main idea of the article are more likely to be of the interest of the readers. To this end, this ranking strategy uses TF-IDF to measure the similarity between articles.

Authority. Similar to the way we calculated authorities in Section IV-B, we use the popularity of an article as evidence of its importance. As previously mentioned, we use the number of incoming links as the popularity measure for an article. These incoming links are created manually by Wikipedia editors. Therefore, is reasonable to assume that these articles are, to some extent, significant to the general public.

[6]see e.g. http://www.useit.com/alertbox/percent-text-read.html

The distribution of authorities follows a power law distribution, where a small number of dominant articles contain the larger part of all incoming links. In order to compensate for this, we applied a logarithmic smoothing function before the proper normalization. In this way we still exploit the information but counterbalance the dominance of the few top authorities.

The output of the HyperLink Ranker is a linear combination of the six contextual hyperlink features here discussed: $TermFrequency$, $Location$, $Length$, $Generality$, $Relatedness$, and $Authority$. We did not apply any weight to the different features. However, in the result section we analyze which features played the most prominent role in our ranking strategy.

V. Evaluation Setup

The goal of our study is to explore the current practices and appreciation of automatically generated links. In Section III we have seen that links on news sites are mainly generated links; manually created links are only rarely inserted by authors of news articles. In this section we describe the setup of a study that explores the nature of generated links and to what extent users appreciate these links in comparison to manual links.

We have chosen Wikipedia articles as the target domain of our study for a variety of reasons. First, as we discussed in the related work (Section II), Wikipedia articles are richly linked; these links are manually created, following strict guidelines. Still, the structure of Wikipedia is very similar to the rest of the Web. These factors make Wikipedia a representative domain. Second, due to the popularity of Wikipedia, our prospective participants are likely to be familiar with its setup and structure, which makes it easier to judge articles. Also, as Wikipedia is an encyclopedia, it provides a solid base for generating links using the methods described in Section IV.

Our dataset consists of a snapshot of the whole Wikipedia corpus from October 2011. It contains more than 4.5 Million pages (all articles without redirect pages). Additionally, we collected the list of Wikipedia categories from the same time period and statistical information of the most linked articles.

For the user evaluation, we sampled the data to create a smaller yet representative set of articles. This sampling was based on the popularity of the pages and chosen to increase the probability of presenting common articles to the user about concepts they are familiar with. We considered articles as popular if they appear within the list of the top 100 most viewed articles within a period of one month. We collected this data from www.wikiroll.com for the period from 1st of February 2011 to 29th of October 2011. This list contained 604 distinct articles from which 100 were randomly chosen.

The average article length of the sampled articles is 42,917 characters or 6,298 words, with an average of 327 outlinks

Table II
OVERLAP OF THE LINKS BETWEEN THE THREE DIFFERENT STRATEGIES.

	L-Recommender	L-Detector	L-Extractor
L-Recommender	-	0.1%	8.7%
L-Detector	0.1%	-	5.0%
L-Extractor	8.7%	5.0%	-

per article and belonging to 18.76 different Wikipedia categories in average. For each of the 100 articles, we applied the three linking strategies proposed on Section IV and generated three distinct lists of links. Table II shows the existing overlap of the lists among the strategies.

For the evaluation, we built an online interface, where each participant was presented with one randomly chosen article from the set of sampled articles at a time. Each article looked exactly as they look in Wikipedia, except that all links were removed. The participants were instructed to quickly read the article to have a general overview of what was the article about.

The participants were then asked if they were familiar with the topic discussed in the article. Regardless whether the answer was positive or negative, the participants were then presented with three distinct lists of links. Each list contained the top ten links as generated by the three linkage strategies described in Section IV. The participants were asked the question: 'Which list provides the best links for the article?'. We also ensured that the lists were randomly assigned to one of the three columns, to ensure that participants would not be inclined to vote, for example, on the left list after a few judgments.

We deliberately did not specify in which sense the links should be 'better'. Imposing any criterion on the 'goodness' of the links would inevitably have introduced a bias towards one of the strategies. Therefore, we left the interpretation of what is 'best' to the participants. The participants were asked to judge the relevance of the suggested links by only looking at the keywords in the lists. Optionally, they were able to click on a link and read the target article for further information. The participants also had the option to skip the evaluation of a given article at any point, in case they did not feel confident to judge the content or the given lists of links. We kindly asked each participant to repeat the evaluation process for at least 10 articles. The average time needed to complete the procedure was estimated to be 10-15 minutes.

The invitation to participate in the study [7] was distributed via several mailinglists and social media. Participation was voluntary and anonymous, and no financial compensation was given.

VI. RESULTS

The main part of this section will be the results of the user study. However, in order to better interpret the results, it is worthwhile to first take a look at the links that were generated or extracted using the three strategies.

[7]available at http://www.l3s.de/~kawase/wikieval/start.php

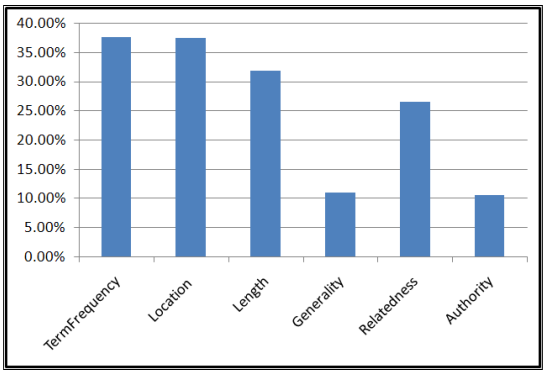

Figure 1. Agreement of each contextual feature with the HyperLink Ranker (the combination of all features).

A. Linking Strategies

In Table II the overlap of the top-10 links generated by each strategy. The low degree of agreement suggests that the relations captured by each strategy tends to be different. In particular the similarity-based L-Recommender and the authority-based L-Detector (with an overlap of 0.1%) generate almost mutually exclusive link sets. The L-Extractor, which generates the top-10 links based on the ranking of manually created links, agrees to some extent with both the L-Detector (5.0%) and the L-Recommender (8.7%) - which suggests that hand-made links contain both similarity-based and authority-based links, but also many 'other' relations.

As explained in Section IV-C, the L-Extractor makes use of a Hyperlink Ranker, which ranks links based on a diverse set of contextual features. None of these features had greater weight than other features. Assuming that the combined evidence from the different features led to an unbiased and more or less optimized ranking, we inspected which features are most prominent in the extracted links. Figure 1 shows the influence (agreement) of each feature with the Hyperlink Ranker (regarding the top 10 results - recall@10).

Term Frequency, *Location* and *Length* turned out to be the most influential features. Each list based on only one of these features agreed for more than 30% with the links based on the combined features. This suggests that these three features capture rather similar information: linked terms that appear often in one article, tend also to appear in the introduction. Further, these linked terms - usually common terms that are familiar to most people - tend to lead to articles with a large average length, which confirms the observations of David et. al [5]. The linked articles tend to be *Related*, in terms of textual similarity.

Interestingly, the influence of the features *Generality* and *Authority* is rather low - even though each feature had the same weight. These features appear to cover different information than the other features - which is in line with the small overlap between the L-Recommender and L-Detector strategies observed earlier. It also suggests a preference of authors to link to similar articles rather than to authoritative

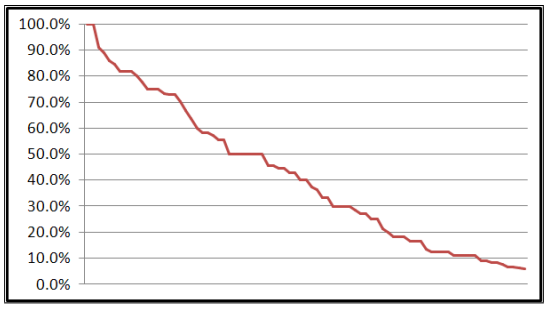

Figure 2. Distribution of participants' familiarity with each article. The line indicates the percentage of participants' who stated to be unfamiliar with the subject - sorted from high to low.

articles.

B. User Study

A total of 102 participants (42 female and 60 male) responded to our invitation that was distributed via various channels. The average age of the participants was 36.2, ranging from 22 to 67 years old. In total, the participants evaluated 972 items, covering all the 100 available articles under evaluation.

Figure 2 shows the familiarity of our participants with the (subject of the) Wikipedia articles. Note that the distribution of familiarity ratings is rather linear, which indicates that the sampled articles concerned topics that the participants were rather familiar with (otherwise, the distribution would have followed a power law). Only three articles were unfamiliar to all participants that evaluated them. The distribution of familiarity also demonstrates the diversity in the participants' background.

In 158 cases the participants skipped a proposed article. In most of the cases (86%), this happened when the participants were not familiar with the subject [8]. Additionally we logged the time each participant took during the evaluation of the articles. In average each participant took 44.93 seconds to evaluate an article and choose a specific list of links. Familiarity with the article had a small, yet significant, impact on the times of the tasks. Tasks that participants stated to be familiar with the article took in average 42.3 seconds while the unfamiliar ones took in average 52.7 seconds.

The results show that the hand-made links from the L-Extractor were preferred in about 51% of the cases. The automatically generated similarity-based links from the L-Recommender was preferred to a slightly lesser extent, but still covered 45% of the cases. Both demonstrate significant users preference ($p < 0.01$) over the least preferred authority-based links provided by the L-Detector, which was chosen in only 4% of the cases. Still, generated links (L-Recommender and L-Detector combined) were considered 'best' in about half of the cases. Given the low overlap between the link sets, this suggests that generated links are a useful complement or alternative to manual links.

[8]During the evaluation the participants were first asked if they were familiar with the article and then given the option to skip.

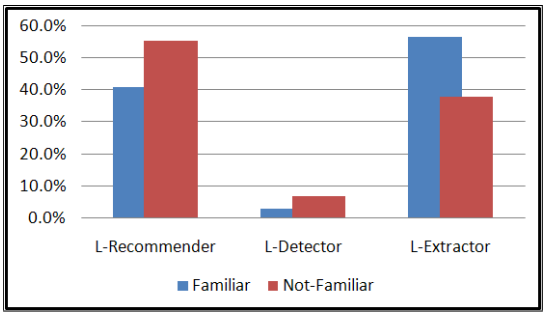

Figure 3. Distribution of the participants' choices for 'best' list of links, split according to the participants' indicated familiarity with the articles' subject.

It is a likely assumption that familiarity with the topic plays a role in the preference for either similarity-based recommendations or manually created links. This assumption is confirmed by Figure 3, in which the preferences for the three strategies is split based on the indicated familiarity with the topic of an article. Note that out of the 972 choices, 62.3% concerned articles of which participants stated to be familiar with.

Figure 3 shows that users who are familiar with a topic, have a clear preference for manually created links (L-Extractor). Conversely, users who are not familiar with the topic, chose the similarity-based links (L-Recommender) in most of the cases. On the one hand this suggests that hand-made links are considered more interesting if one already has some background knowledge on a topic. On the other hand, it also suggests that similarity-based links and keywords - which are textually closer to the article - are considered as useful starting points for those who are not familiar with the topic.

VII. DISCUSSION

The results from our user study do not reveal a strong preference for either one of the categories - manual links were preferred to automated links in only about 50% of the cases. This suggests that for many user contexts, generated links and recommendations are considered *good enough*. As a consequence, it may be a defendable choice for a site to rely on algorithms for generating links in order to reduce human effort.

A rather controversial observation is that similarity-based automated links worked better than hand-made links for users who are not familiar with a topic. It would be too speculative to draw any definite conclusions from this observation. It may be the case that similar articles are a better starting point for familiarizing with a topic [13] than human-created links, which are assumed to be less straightforward, more serendipitous and more associative. It may also be the case that the *link terms* as created by human authors provide less information scent [10] to users and therefore are

less effective in communicating the relevance of the content behind the link. In any case, this is a topic that would need more investigation.

The results of the user study also indicate that links to authoritative pages - which typically provide background information on a person, place or other entity - are considered less useful than other types of links. As we have seen, links to entities are one of the most common types of links on news sites - a practice that has received mixed responses (as discussed in the related work). The low preference for the authoritative links provided by the L-Detector is in line with these mixed responses and is yet another motivation for being conservative with linking 'everything to everything'.

Despite the increasing prevalence and perceived usefulness of automated links, our participants preferred manually created links in situations when they were already familiar with a topic. Furthermore, the small overlap with the set of generated links indicates that hand-made links indeed provide additional, complementary value. Therefore, it would be sensible not to throw out manually created hyperlinks altogether: even though links to similar articles and authoritative pages could and perhaps should be automated, there is still the need for links 'suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain' [4].

VIII. Conclusion

In this paper we investigated to what extent automatically generated hyperlinks may provide an alternative for manually created links. Automated links are already very common on the Web, as we have shown with an analysis of current linking practices on news Web sites. By contrast, links within Wikipedia are still manually (and collaboratively) created, according to strict guidelines.

First, based on the proposed *Hypersonomy* model to exploit contextual information of hyperlink assignment, we developed the Hyperlink Ranker that generates a ranked list of the most important hand-made links of a page (Link Extractor). Then, in a user study, we compared hand-made links within Wikipedia with automated links that were based on either similarity or authority. The results show that user preference for authority-based links - which usually lead to reference pages - is rather low. By contrast, similarity-based automated links to related articles were almost as successful as hand-made links (in terms of user preference). In situations where users were not familiar with a topic, similarity-based links were even preferred.

The success of automated links does not automatically imply that manually created links should be abandoned. On the contrary, hand-made links are particularly useful for creating more serendipitous, associative connections that are of benefit to users who are already (slightly) familiar with a topic. However, co-existence of these different breeds of connections raises the question how to inform users about how and why a link was created. To a certain extent this is already achieved by labeling navigation boxes with terms as 'latest articles' or 'related content'. But this solution does not work for within-text links. Is it necessary for users to know whether a link to an entity is automatically generated or not? The mixed responses to such automated links on news sites suggests that the answer is yes and that suitable visual metaphors need to be designed and accepted by the community.

The benefits of automated links are evident: they enrich the hyperspace with more and new connections; they also relief Web authors from the burden of finding similar content and creating links themselves. The division of work is not yet clear. Future research should find more specific pointers to what extent and for which situations algorithms can satisfactorily replace human authors.

References

[1] S. F. Adafre and M. de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, LinkKDD '05, pages 90–97, New York, NY, USA, 2005. ACM.

[2] J. E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1095–1096, New York, NY, USA, 2008. ACM.

[3] P. Brusilovsky. *Adaptive Navigation Support*, pages 263–290. 2007.

[4] V. Bush. As We May Think. *Atlantic Monthly*, 176(1):641–649, March 1945.

[5] C. David, L. Giroux, S. Bertrand-Gastaldy, and D. Lanteigne. *Indexing as problem solving: a cognitive approach to consistency*, volume 32, pages 49–55. 1995.

[6] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM.

[7] D. Milne and I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *In Proceedings of AAAI 2008*, 2008.

[8] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[9] T. Nelson. Xanalogical structure, needed now more than ever: Parallel documents, deep links to content, deep versioning, and deep re-use. *ACM Computing Surveys*, 37 (4es), 2000.

[10] P. Pirolli and S. K. Card. Information foraging. *Psychological Review*, 106:643–675, 1999.

[11] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[12] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975.

[13] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 415–422, New York, NY, USA, 2004. ACM.