Automatically Embedding Newsworthy Links to Articles: From Implementation to Evaluation

Ioannis Arapakis and Mounia Lalmas
Yahoo! Research Barcelona, Avinguda Diagonal 177, Barcelona, Spain 08018.
E-mail: arapakis@yahoo-inc.com; mounia@acm.org

Hakan Ceylan
Netflix, Inc., 100 Winchester Circle, Los Gatos, CA 95032. E-mail: hceylan@netflix.com

Pinar Donmez
Banjo Inc., 250 King Street #772, San Francisco, CA 94107. E-mail: pdonmez@gmail.com

News portals are a popular destination for web users. News providers are therefore interested in attaining higher visitor rates and promoting greater engagement with their content. One aspect of engagement deals with keeping users on site longer by allowing them to have enhanced click-through experiences. News portals have invested in ways to embed links within news stories but so far these links have been curated by news editors. Given the manual effort involved, the use of such links is limited to a small scale. In this article, we evaluate a system-based approach that detects newsworthy events in a news article and locates other articles related to these events. Our system does not rely on resources like Wikipedia to identify events, and it was designed to be domain independent. A rigorous evaluation, using Amazon's Mechanical Turk, was performed to assess the system-embedded links against the manually-curated ones. Our findings reveal that our system's performance is comparable with that of professional editors, and that users find the automatically generated highlights interesting and the associated articles worthy of reading. Our evaluation also provides quantitative and qualitative insights into the curation of links, from the perspective of users and professional editors.

Introduction

In today's world, news portals have become a popular destination for millions of web users around the globe who wish to stay informed about current events, local and global (Purcell, Rainie, Mitchell, Rosenstiel, & Olmstead, 2010). Because there is great potential for online news consumption but also serious competition among news portals, providers are constantly investigating effective and efficient strategies to engage users longer on their sites. User engagement (Attfield, Kazai, Lalmas, & Piwowarski, 2011), the *emotional, cognitive, and behavioral connection that exists between a user and a resource*, is the result of trustworthy, quality, relevant, and entertaining content. In this article, we evaluate one type of strategy-promoting engagement: *enticing users to browse a site through embedded links within news articles*.

Embedded links, or hyperlinks, are an effective strategy for prolonging time users spend on a site by sustaining engagement through interactive click-through experiences, encouraging inquisitive behavior, and providing a feeling of richness and control (Attfield et al., 2011). Figure 1 displays an example of a news article from Yahoo! News, where the embedded links highlight interesting or important pieces of information aiming to attract users' attention. Once clicked, the links redirect users to another page showing the referenced content. We studied this phenomenon in the context of news portals, where the embedded links direct users to other pages within the same domain. Users reach the information that is just one click away and, it is hoped, become more engaged with the site, and stay longer by reading more news stories. However, hyperlinks are mostly created by human editors, making it a manual task that is time consuming and not scalable.

Received October 26, 2012; revised January 23, 2013; accepted February 8, 2013

 DOI: 10.1002/asi.22959

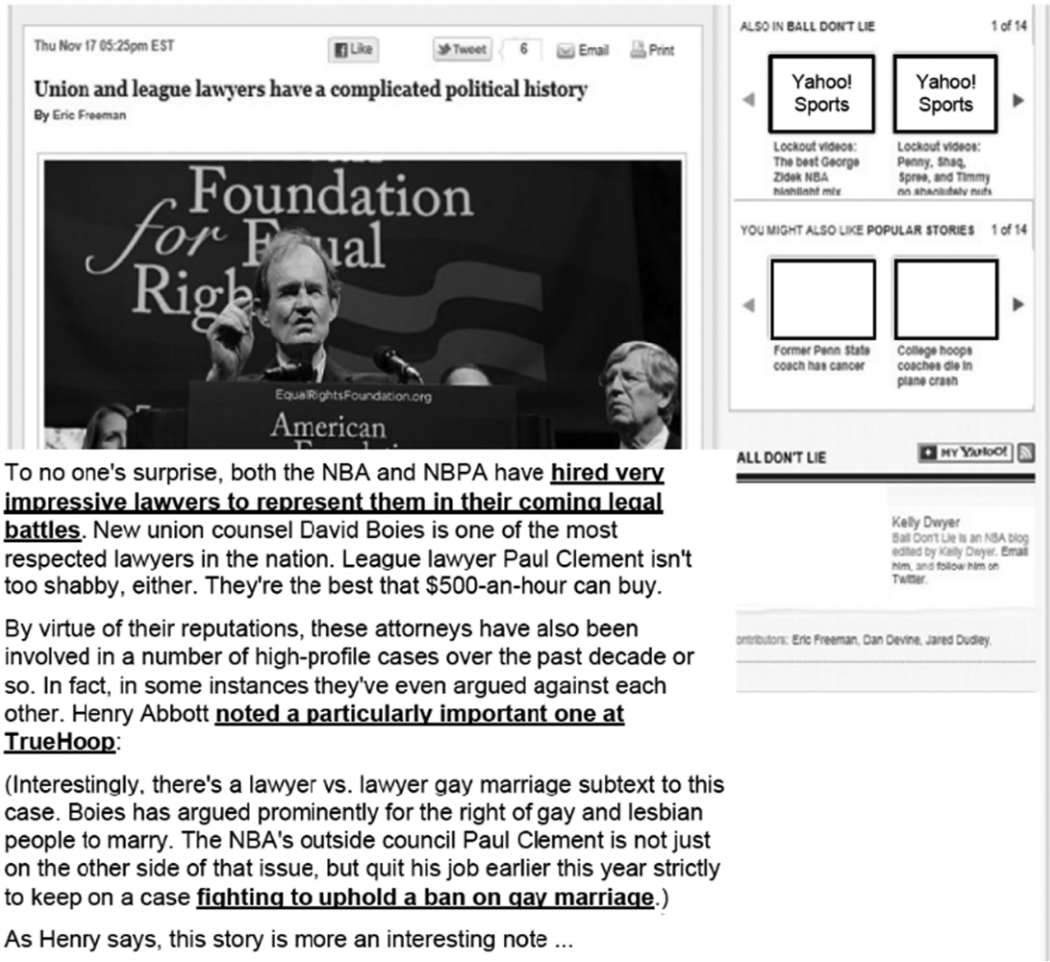

FIG. 1. A news article with embedded links highlighting the newsworthy events selected by professional editors.

In this article, we propose an automatic approach to hyperlinking, where for any given news article the goal is to identify newsworthy events as a potential source for links. Newsworthy events are more likely to have a related news article already written in the past. For example, in the sentence "The Boston Celtics star announced his retirement from professional basketball yesterday by tweeting a link to a 10-second video," the phrase "The Boston Celtics star announced his retirement," a *newsworthy event*, would be highlighted and linked to the corresponding article devoted to the announcement. Our system had two main requirements: First, it could not rely on resources such as Wikipedia to identify events, because it is not likely that actual newsworthy events (such as the one mentioned earlier) are, or will ever be, Wikipedia concepts. Second, our system should be mostly domain independent because our long-term goal was for it to be deployed across several domains (news, entertainment, finance, etc.). We ensured these by using established text processing and understanding techniques, which we adapted to fit our context.

Because of its relative simplicity, it was important to carry out a comprehensive evaluation of the proposed system. To this end, a systematic and rigorous evaluation was performed to assess the system-embedded links against manually curated ones by professional editors and to understand the effects of embedded links from a user's perspective, focusing on the associated reading experience. Carefully designed experiments, harnessed by the crowdsourcing power of Amazon's Mechanical Turk (MTurk), provided several quantitative and qualitative findings that helped us better understand the curation of links from the perspective of users and professional news editors, and offered empirical support to the *why* and *how* associated with the resulting news reading experiences. The results indicate that our automatic approach to link curation offers a news reading experience that is comparable with that of manually curated links, delivered by professional editors. Because the manual creation of hyperlinks is a time-demanding and challenging task, the proposed system, with its massive scalability and domain independence, has the potential to reduce the manual labor required for this task and improve the efficiency of the editors.

This article expands on previous work presented by Hakan, Arapakis, Donmez, and Lalmas (2012), and focuses

DOI: 10.1002/asi

on assessing the quality of the system-embedded links and the resultant reading experience. It provides significant additional contributions, such as the presentation and analysis of two further experimental studies (a pilot study and a study of the reading experience), a qualitative investigation of the effects of embedded links, and a correlation analysis of several experimental variables. Finally, the thematic analysis of open-ended questions sheds further light and provides empirical support to our findings. The rest of the article is organized as follows: The Related Work section reviews related work. The System section presents the proposed system in detail, whereas the Experiments & Results section reviews the experiments and the results. We discuss our main findings in the Discussion section and conclude in the Conclusions section.

Related Work

Work in automatically generating hyperlinks existed already in the 1990s (Allan, 1995; Bluestein, 1999; Ellis et al., 1996; Green, 1997, 1998). It has recently attracted fresh interest, primarily within the context of Wikipedia, looking at cross-referencing documents (Knoth, Zilka, & Zdrahal, 2011; Milne & Witten, 2008; Wang, Li, Wang, & Tang, 2012), missing link detection (Fissaha Adafre & de Rijke, 2005), automatic approaches for hyperlinking (Gardner & Xiong, 2009; He & de Rijke, 2010), automatically linking documents to encyclopedic knowledge (He, de Rijke, Sevenster, van Ommering, & Qian, 2011; Mihalcea & Csomai, 2007; Shen, Wang, Luo, & Wang, 2012), and entity linking augmented with human intelligence (Demartini, Difallah, & Cudré-Mauroux, 2012). Link generation approaches are also used in disambiguation tasks (Bron, Huurnink, & de Rijke, 2011; Cucerzan, 2007; Jijkoun, Khalid, Marx, & de Rijke, 2008). Many of these works are not directly comparable with ours because of different aims and a reliance on properties specific to Wikipedia. Our focus is to identify the newsworthy events in a given news article and connect them to the appropriate news articles, where it is unlikely that all newsworthy events are Wikipedia concepts. We therefore used, instead, well-known text processing and understanding techniques, which we adapted to fit our context.

The Initiative for the Evaluation of XML (INEX) retrieval launched the Link-the-Wiki track in 2007, the goal of which was to automatically generate links among Wikipedia pages (Huang, Geva, & Trotman, 2009). Techniques aimed mostly at identifying highly similar pages to a given page, from which candidate links that might be missing on the given page were identified. Applying the same techniques in our context would connect news articles that are similar in terms of content, but not the newsworthiness of events as investigated in this article.

The works most closely related to ours are those of Mihalcea and Csomai (2007), Milne and Witten (2008), He et al. (2011), and Shen et al. (2012). The Wikify! system (Mihalcea & Csomai, 2007) uses Wikipedia for automatic keyword extraction and word sense disambiguation. These two tasks are combined to identify the important concepts in the text and link them to the corresponding Wikipedia pages. Given a document, their system automatically extracts important words and phrases in the document, and identifies for each of them the appropriate link in the Wikipedia repository. Results showed that the manual (performed by human participants) and automatic annotations were not distinguishable. However, Wikify! is heavily tailored for Wikipedia, where the most important concepts have their own corresponding Wikipedia pages. This is unlikely to be the case for newsworthy events.

Similar work (Milne & Witten, 2008) attempted to enrich Wikipedia and other collections with links, using machine learning techniques to distinguish link-worthy topics from others. Contextual information in the source text was used to determine the best related Wikipedia concepts, which, in turn, also served as features for anchor text detection. Amazon's MTurk was used for the evaluation, where the workers judged the validity of the embedded links and provided missing links, if any. We were inspired by their evaluation method, and designed two large experiments using Amazon's MTurk to assess both the quality of the system-generated links and their effect on the news reading experience.

The work presented by He et al. (2011) attempted to generate links from medical reports to Wikipedia pages for explanations or background information. They showed that the approaches by Mihalcea and Csomai (2007) and Milne and Witten (2008) did not yield satisfactory results because medical phrases typically have a more complex syntactic and semantic structure than Wikipedia concepts. They, therefore, developed their own approaches. Events, as phrased in a news article, also have a specific structure (discussed in the System section); in addition, very few form concepts in Wikipedia. Similarly, we had to design our own automatic link generation approach.

The work discussed by Shen et al. (2012) addresses the challenge of semantic disambiguation in an entity-linking task. The authors proposed an approach that uses a knowledge base unifying Wikipedia and WordNet, by leveraging the rich semantic knowledge derived from Wikipedia and the taxonomy of a knowledge base (YAGO) (Suchanek, Kasneci, & Weikum, 2012). An empirical evaluation of the framework was carried out using two publicly available data sets annotated for that purpose. However, the performance of the proposed framework was evaluated and compared against the other systems solely in terms of its linking detection capability (accuracy). One missing aspect is the involvement of human evaluators, as well as a qualitative assessment of the generated links and the associated reading experience. In this article, we pay particular attention to the evaluation of the proposed system from both a user perspective and a professional editor perspective.

Our work is related to news recommendation systems (Agarwal, Chen, & Elango, 2009; Lv, Moon, Kolari, Zheng, Wang, & Chang, 2011) whose goals are to present users with

DOI: 10.1002/asi

related articles, given the article they are reading. These systems can be personalized using signals extracted from users' past behavior. Contrary to our system, these approaches process entire articles rather than specific subsections. For example, an article on the "2012 US Elections" could lead to the recommendation of an article on the "2012 Elections in China." This is an unlikely scenario for our system unless the Chinese elections are explicitly mentioned in the first article. For that first article, our system could detect the event "Barack Obama announced his candidacy for presidency . . ." and refer the user to a corresponding article via a hyperlink in the sentence. Depending on its implementation, the same article may or may not be suggested by a recommendation system. In general, news recommendation systems are interested in broader concepts, whereas our system handles specific mentions of events in the article. Nevertheless, the ultimate goal for both is to promote user engagement on the site, although this result can be accomplished in many different ways.

A main challenge in improving automatic link generation systems is the difficulty to evaluate the results. The issue that makes reliable and systematic evaluation problematic is related to both technical and cognitive aspects. The difficulty in obtaining the ground truth for a sufficiently large data set is caused both by the lack of human resources and the inherent subjectivity of the task. Currently, the evaluation approaches used in existing work vary between a statistical examination of entire collections and relevance judgments provided by a limited number of human judges, whereas in some cases, collections like INEX Wikipedia are used for training and testing purposes. In most system-oriented evaluations, some similarity measure (e.g., based on graph-theoretic techniques) or standard information retrieval metrics (e.g., precision, recall, and accuracy) are applied. Another approach with more ecological validity is bucket testing on a random live set of users, and for a short time interval, usually in combination with interaction metrics such as click-through rate. In few cases, human evaluation of links through voting or selection is used to build the ground truth. However, the feedback provided is limited to a simple voting for characterizing a link that includes answers such as "correct," "incorrect," or "don't know."

Our experimental approach expands on the method proposed by Milne and Witten (2008) by inquiring about the qualitative aspects of link generation, from the viewpoint of both professional editors and end users. We include standard performance metrics like precision, recall, and *F*-score, for comparability purposes, but also develop an in-depth understanding of user behavior and the qualitative aspects that shape the associated reading experience. Also, as reported in the Data Set section, only a small percentage of links was embedded in common by the system and the professional editors; we therefore investigate whether the system links are worthy or relevant. Finally, we try to establish the real-life effect of different system configuration settings on the users and what the results mean. An "optimal" performance indicates nothing if the reading experience is unsatisfactory or if the proposed service fails to satisfy the actual user needs, whereas understanding how people perceive the different qualities of system and editor links provides a ground truth that goes beyond standard metrics.

System

This section describes our system, Linker for Events to Past Articles (LEPA). LEPA aims to identify the newsworthy events in an article and create hyperlinks to their previously published content with the aim to provide users a chance to read more about that event. Because of the nature of the embedded links, LEPA cannot rely on resources such as Wikipedia to identify newsworthy events. In addition, to allow for its potential deployment across different domains, we used known text processing and understanding techniques. LEAP has two main components, an indexer and a linker, each of which are described next.

Indexer

The indexer processes articles over a given time period by extracting features from each article and storing them to facilitate faster retrieval. The indexer runs in two stages: The build stage produces an index for a set of articles over the time period, whereas the update stage is run periodically to add fresh articles to the existing index. For the build stage, we constructed an index from articles that spanned more than a month, whereas for the update stage, we processed new documents daily and added them to the index. In both stages, we implemented a simple inverted index approach. The inverted index stores a list of the documents for each word in the vocabulary derived from the corpus, which is formed from the entire set of news articles being indexed. The frequency of each word in the document is stored in the inverted index. These frequencies are calculated during the feature extraction step of the indexer.

The retrieval task is to find the article in the index that exactly discusses the corresponding event. Finding a precise matching between the article and the event can be more easily accomplished if both contexts have comparable sizes. Because an event consists of only a few words, only the title and the abstract sections of the news articles are considered and indexed.[1]

Linker

The task of the linker is to find newsworthy events in each article and link them to the previously indexed articles. It first identifies sentences that mention newsworthy events, then for each event matches and retrieves newsworthy articles. Finally, the top-ranked article is hyperlinked to the event if it satisfies a certain confidence level criterion. The algorithm is shown in Algorithm 1.

[1]The abstract section corresponds to the first paragraph of a news article.

DOI: 10.1002/asi

ALGORITHM 1 The linking algorithm.

```
Require: ϕ: predetermined confidence level threshold
   for all sentences S in the text do
      if S has no named entities then
         Continue
      end if
      for all verbs V in S do
         if V is in past tense and V is an action verb then
            if V has immediate noun phrases (NPS) then
               Add E = (V, NPS) to the list of events
            end if
         end if
      end for
   end for
   for all events E in the list do
      Construct query Q from E
      Query the retrieval index with Q
      Retrieve top result D with score S
      if S ≥ ϕ then
         Hyperlink E with D
      end if
   end for
```

Selecting the candidate sentences. We identified three important criteria in selecting the candidate sentences, based on an empirical observation of the links curated by professional editors using various news sources: (a) The sentence must contain a named entity, (b) the sentence must contain a verb in past tense, and (c) the verb mentioned in the second criteria must be an *action verb*.

Almost all important events we observed have to do with one or more important entities that occur as the subject or the object of a sentence. For example, the sentences "*Barack Obama* announced his candidacy for presidency on Feb. 10," or "A few days ago *Google* announced their acquisition of *Zagat*, the popular publisher of restaurant review guides" contain named entities *Barack Obama*, *Google*, and *Zagat* as part of the events that refer to news in the past. This criterion could be restrictive by ignoring otherwise good candidates such as "The company issued a press letter yesterday regarding the new privacy policies," where, for example, "*The company*" is a co-reference, because it refers to a named entity such as Google. In this article, we do not use a co-reference resolution approach because of its complexity and the additional noise it might introduce to the system.

The second criterion is trivial. Because our goal is to hyperlink the events in the current article to previously published content, we ensure that the candidate sentence contains a verb in past tense. Our last criterion stems from the need to eliminate verbs that usually do not specify any event, and thus cannot be linked to any previously published content. Examples include *be*, *become*, *seem*, *grow*, among others. We are looking for verbs that describe an action, which are referred to as action verbs. Therefore, we ensure that all identified past tense verbs are action verbs; otherwise, the sentence is eliminated.

The sentences are processed using the Natural Language Processing Toolkit (NLTK) (Bird, Klein, & Loper, 2009), a freely available application for research purposes, to filter the sentences based on these criteria.[2]

Constructing the query. Once the candidate sentences are identified, we extract the events from each candidate sentence. An event is contained in a sentence and is determined by a predicate of that sentence. Similar to first-order logic, we use the term *predicate* to describe a function over arguments. The function is formed by the verb, and its arguments are the noun phrases that are immediately before and after the verb.[3] Thus, it is possible for one sentence to contain more than one event. For instance, the predicate formed from the sentence "Barack Obama announced his candidacy for presidency on Feb. 10" would be "*announced (Barack Obama, his candidacy for presidency)*" as the verb *announce* forms the function, and the immediate noun phrases "*Barack Obama*" and "*his candidacy for presidency*" form the arguments of the predicate. Hence, the event extracted in this example would be "*Barack Obama announced his candidacy for presidency*."

The general pattern being identified is subject-verb-object relationships. This has certain disadvantages, because

[2]NLTK (www.nltk.org) provides utilities to extract named entities, part of speech tags, and so on.

[3]We control the notion of *immediate* via a window parameter defined in terms of number of characters between the verb and the noun phrases.

DOI: 10.1002/asi

it assumes that the verbs are normal transitive verbs that take a single direct object. This is not always the case. A verb can be intransitive, that is, it takes no objects, or it could be transitive but take both a direct and an indirect object, as in the case of *complex transitive* verbs and *ditransitive* verbs *(datives)*. Nevertheless, this leads to an approach that is easily scalable to other languages.

We use NLTK's built-in noun phrase chunker to automatically identify the noun phrases in the sentence, and the verb is identified through the part of speech tags as mentioned previously. Once the event is identified, we use NLTK again to remove the stop words and stem each word in the event, including the verb. The resulting phrase forms our query. Our query for the earlier example would be "*Barack Obama announce candidacy presidency*."

Ranking the results. Once the query is formed, the matching articles in the index are retrieved and ranked. The inverted index, described in the Indexer section, keeps track of each word and the document it appears in together with its frequency. We form vectors of term frequencies, $\vec{q}$ and $\vec{d}_i$, for each query q and document d_i in the corpus, respectively. The dot product of the query vector with a document vector gives us the importance score of that document for the query. We normalize the dot product with the length of the document. Thus, the score of document d_i for the query q is $\vec{q} \cdot \vec{d}_i / |\vec{d}_i|$. The documents are ranked according to this score.

When constructing the vectors for the documents, we give more weight to the frequency of a word appearing in the title of the document. The reason is the more query terms we find in the title, the more confident we are of that document describing the event. For example, consider the event "A magnitude-8.8 earthquake hit Chile" and two matching documents with titles "8.8-Magnitude Quake Hits Chile," and "Millions are Displaced After the Chile Quake." Even though both documents are related to the event, the former matching document is devoted to the event, whereas the latter has only tangential relevance. We set the weight of matching terms in the title to 3, by tuning this parameter on a separate validation set.

Finally, if the score of the top result retrieved from the index is above a predefined threshold, the event is linked to the article in the index. This threshold parameter can be set depending on the application. For example, in a setting where precision is more important than recall, one can set the threshold to a high value to make the system very precise, whereas trading off recall and vice versa. The thresholds experimented with in this article are listed in the Design section.

This ends the description to LEPA, a system that automatically embeds newsworthy links to articles. Our system combines existing domain-independent methods and tools from the text processing and understanding area, and does not require any training data or human intervention, making it an attractive and generalizable solution to many domains. The next step is to evaluate the generated links.

Experiments and Results

In this section, we present a series of experiments assessing the benefits of LEPA. We aim to: (a) understand better the curation of links for news content from the perspective of professional editors and users, (b) measure the performance of the proposed system using standard metrics, (c) compare the quality of system-embedded against manually curated links, and (d) evaluate the overall news reading experience.

Data Set

We used a collection of 200 news articles taken from the top 50 most viewed articles of Yahoo! News on four different dates. To mitigate any unwanted effects stemming from factors such as document length or topicality, we kept our selection random, covering a diversity of topics and a range of document lengths (articles varied between 150 and 2,000 words). The news articles were separately annotated by LEPA and a team of professional editors from the same online news provider. The articles were available in *simple HTML* format (including content of the article with embedded links) and in *rich HTML* format (including the original crawled content with embedded links). We repeated this process for our system using four different precision settings as discussed in the Design section.

The editors, who had no prior knowledge that their work would be evaluated against the proposed system, were asked to read the articles, and identify events and entities that were good candidate links. The guidelines indicated this as a routine editing task and instructed them to link articles that were perceived as related and newsworthy, and that would provide interesting insights with respect to the main article. The only limitation was that the linked articles had to reside within that news provider's site. The editors were allowed to embed as many links as they thought appropriate.

Of the 200 articles, we retained 75 after filtering out the articles for which the system did not detect any events. Our system identified a total of 192 links, whereas the editors identified 211 links. We excluded 28 links that were embedded in common by LEPA and the editors, as our focus was to compare the quality of system-embedded against the manually curated links. As common links we treated those cases of anchored text that appeared in the same article, same paragraph/sentence, and shared at least one common word. This resulted in 164 system-embedded and 183 manually curated links. From the latter, we retained a random selection of 164, to have an equal contribution of both types of link.[4] Because the common cases consisted of only a small fraction (7%) of the collection, we regard this as an indication that the system links were complementary to those of the professional editors. It is therefore important to evaluate the system-generated links, not only in terms of their quality

[4]In this case, best practice would be to apply sampling and present the average results. However, a cross validation was not feasible in our case because of the limited resources.

DOI: 10.1002/asi

as assessed by (Pilot Study section) or compared with those generated by (Assessing the Links section) professional editors, but also, and in particular, with respect to their overall effect on the news reading experience (Assessing the Reading Experience section).

MTurk

For our online experiment, we used the Amazon MTurk crowd-sourcing service. According to Mason and Suri (2010), MTurk combines several benefits for running online experiments. First, it offers access to a large pool of candidate participants with a fairly stable availability over time. Another advantage is the diversity of workers' background, in terms of age, ethnicity, and socioeconomic status. Finally, the low cost at which studies can be conducted makes it an attractive solution compared with the more costly laboratory experiments. MTurk served as the means to conduct a large-scale, labor-intensive study under strict time constraints, although we had to account for several limitations that are common to online experimentation, such as threats to ecological validity, lack of control over the experimental setting, distractions in the physical environment, anonymity of participants, among others. While considering these, we took preventive measures to discount low-quality responses and undesirable participants by using validation tests and strict selection criteria.

Pilot Study

A preliminary evaluation of the system-embedded links was performed to understand better the curation of links from the perspective of professional news editors. For this assessment, we created 164 tasks, one for every combination of article-link. The links were generated using the system with precision threshold 0.0 (see Design section). For each task, the editors (employees from Yahoo! News not involved in the annotation task discussed in the Data Set section) were given a news article in simple HTML format that contained one embedded link as shown in Figure 2. We intentionally opted for this particular layout to reduce any unwanted effects and cognitive biases caused by the visual saliency of nonrelevant elements (images, ads, text) of the original crawled pages. The content of each link was presented in a pop-up box, allowing for a quick and easy evaluation.

The editors rated the links using a 4-point Likert scale commonly used in their evaluations: (1) *bad*, (2) *fair*, (3) *good*, and (4) *excellent*. Only the first option indicates that the link was detected incorrectly. Options 2 to 4 suggest an average to excellent link quality. The results of this assessment reveal that, within the collection ($N = 164$) of system-embedded links, 35.15% of them were regarded as bad, 34.93% as fair, 20.83% as good, and 9.09% as excellent. Taking into account that this evaluation was carried out by

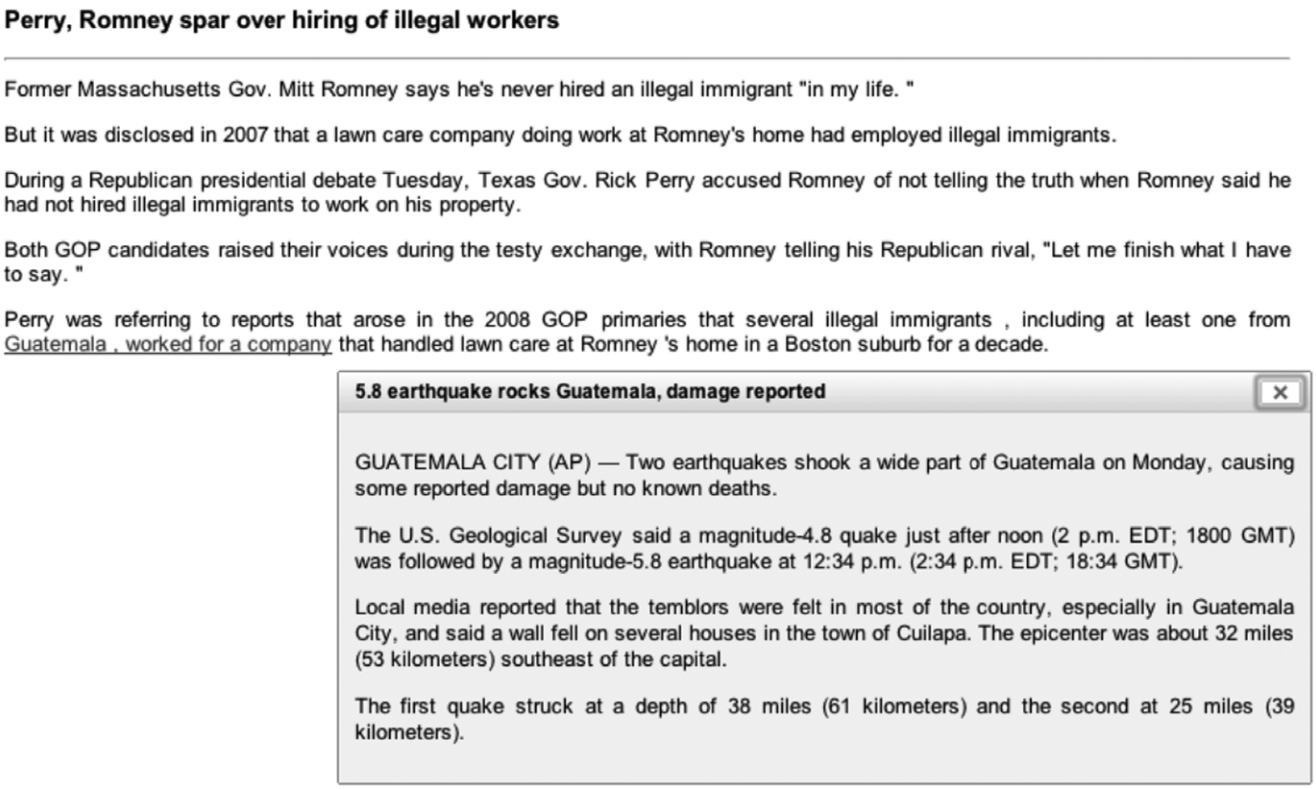

FIG. 2. Example of news article with an automatically augmented link to a related article.

DOI: 10.1002/asi

professional editors with high standards in news editing, we decided to regard the fair links as acceptable. With 64.85% of the links being "good enough," this first study provides initial evidence of LEPA's potential to support the link curation process, making the task less time consuming and laborious for the editors.

Assessing the Links

We examined more closely the performance of our automated approach using standard metrics, quantitative and qualitative judgments from human evaluators, and compared the system-embedded links against manually curated ones.

Design. This study used a between-groups design ("group A," "group B") with three independent variables: type of link (two levels: "system-embedded," "manually curated"), precision configuration (four levels: 0.0, 0.1, 0.2, 0.3) and date of publication (four levels: "19/10/2011," "20/10/2011," "16/11/2011," "17/11/2011"). The type of link was controlled by introducing either system-embedded (group A) or manually curated (group B) links. The precision configuration was controlled by adjusting accordingly a threshold value in LEPA (see Ranking the Results section). The results were filtered based on this confidence level criterion, and those that did not make the cutoff were dropped. This allowed some control over the system's levels of precision and recall, producing results that varied between *high precision-low recall* and *low precision-high recall*. The date of publication was controlled by constructing our experimental data set with news articles crawled on four different dates, thus reducing the dependency of our findings on the temporal factor.

Each participant took part in one condition (one article-link combination) and assessed three different aspects of the task: (a) the main article, (b) the associated article, and (c) the link. We also measured the system's performance. Participants assigned to group A evaluated the system-embedded links, whereas participants from group B evaluated the manually curated ones. With respect to the first category, the dependent variables were: (a) interest, (b) newsworthiness, and (c) similarity to other news read online. In terms of the second category, the dependent variables were: (a) type of relation with the main article (five levels: "related to the main topic of the article," "related to a subtopic of the article," "tangentially related," "unrelated," "other"), (b) newsworthiness, and (c) interesting insights with respect to the main article. Regarding the third category, the dependent variables were: (a) suitability of the anchored text,[5] and (b) relatedness with the associated article. Lastly, the system performance was measured using the standard metrics of precision, recall, and *F*-measure.

Tasks. We prepared 328 tasks, each a unique combination of article-link, using the 75 news articles and corresponding 164 + 164 links discussed in the Data Set section. Each participant was assigned a single, randomly selected article (Figure 2). While reading the article, the participants were instructed to click on the link that appeared in the text and go through the associated article that it pointed to. The articles were presented in simple HTML format to mitigate any unwanted effects stemming from the visual saliency of nonrelevant elements of the original content. Each task was performed by two different participants to reduce the subjectivity of individual responses. Upon completing the task, the participants were redirected to an online questionnaire.

Questionnaire. A post-task questionnaire was used to elicit information on several aspects of the task such as the main and associated articles, the quality of the embedded links, and the participants' reading experience. A demographics section gathered background information and inquired about previous experience with online news reading. All questions were forced-choice type using a 5-point Likert scale. Questions asking for user rating on a unipolar dimension have the positive concept corresponding to the value of 5 and the negative concept corresponding to the value of 1. We furthermore introduced follow-up questions, conditional upon each response, to elicit the reasons behind the participants' choices.

Procedure. We designed the task and the questionnaire in such a way that completing them accurately and in good faith required approximately the same amount of effort as random or malicious completion. To discourage cross-site scripting attacks and ensure that the tasks were performed by human participants, we applied input validation twice for each task. Input validation is a method that can significantly increase the quality of the obtained data when they include questions that have verifiable answers (Kittur, Suh, Pendleton, & Chi, 2007). For the first validation, we used Captcha[6] in the introductory page of the tasks. For the second validation, we used keyword tagging with respect to the theme of the main and associated articles. Keyword tagging is a form of verifiable question that can raise the cost of generating invalid, malicious responses and increase time on task. In addition, we recorded the times required to complete each step of the task. This allowed us to distinguish automated responders from human participants. A final check was to accept as participants only workers who had gained a high reputation from other requestors, by having at least 90% of their responses to previous tasks accepted, as well as a number of completed human intelligence tasks (HITs) greater than or equal to 50. The participants were asked to complete the task, including the questionnaire, in a single sitting. They were also informed of their option to opt out from the task at any point without being compensated. The payment for participation was $0.66.

[5]As anchored text, we refer to the underlined text (sentence, phrase, etc.) that appeared in each hyperlink.

[6]www.captcha.net

DOI: 10.1002/asi

Participants. Six hundred and sixty-four participants were recruited through MTurk, from which we reached the expected number (328×2) of approved assignments from 656 different participants, who spent an average of 17.68 minutes on each task and provided a total of 195 hours of labor. The participants were randomly distributed into two even groups (groups A and B), and were of mixed ethnicity and educational background. Group A consisted of 54.87% male and 45.12% female participants mainly younger than 41 years (83.53%), with the largest group between the ages of 24 and 29 (27.5%). They were all proficient with the English language, and the majority (87.49%) indicated reading the news online between a few times per week to several times a day. Group B consisted of 61.3% male and 38.69% female participants mainly younger than 53 years, with the largest group between the ages of 24 and 29 (31.84%). They were all proficient with the English language, and the majority (85.41%) indicated reading the news online between a few times per week to several times a day. The Mann–Whitney test did not indicate any statistically significant difference between the two groups in terms of age, gender, educational level, proficiency with English, or frequency of reading the news.

Results. We evaluated the performance of LEPA against professional editors on the selected 75 news articles, with a total of 328 links. The performance was measured using macro-average precision, recall, and *F*-measure, and as ground truth for relevance, we used the participants' assessments. Per article, precision was computed as the number of correctly embedded relevant links (links that received in terms of relatedness a score ≥ 3 on a 5-point Likert scale) over the number of all possible embedded links for that article (the links that were embedded by the system with the lowest threshold of 0.0). Similarly, recall was computed as the number of correctly embedded relevant links in each article, over the number of all possible relevant links for that article. The macro-average *F*-score was computed as the harmonic mean of these two figures.

The Mann–Whitney test, the chi-square goodness-of-fit, and the chi-square test of association were used to establish the statistical significance ($p < .05$) of the differences observed in the experimental results, as well as isolate the significant pair(s) through pairwise comparisons. To deal with Type I errors, we applied a Bonferroni correction, and so all effects are reported at a .005 level of significance.

Quantitative findings on news articles and embedded links. To evaluate the main article, we asked our participants to provide scores for the following questions: (a) "Did you find the article informative?" (b) "How interesting did you find the article you read?" and (c) "Was the article you read similar to other news you usually read?" In this evaluation, the results are presented across all groups because we were interested in examining the overall effect of our experimental manipulation on our sample, instead of narrowing it down to specific subgroups. For the first question, the participants reported the main article as somewhat to very informative (mean $[M] = 3.5914$, standard deviation $[SD] = 0.6164$). Regarding the second question, the participants felt that the main article was somewhat to very interesting ($M = 3.3978$, $SD = 0.6825$), whereas in the third question, they rated it as somewhat similar to other news that they usually read ($M = 2.8841$, $SD = 0.7832$). These findings indicate that the news articles we used were a fair approximation of what users read online. Moreover, the scores assigned to *interest* and *informativeness* suggest that these variables did not suffer from any adverse effects introduced by the manipulation of the independent variables.

Table 1 presents the frequency scores for all five types of links, in relation to the question: "Please indicate if the associated article is related to the overall theme of the main article, related to a subtopic within the main article, tangentially related or unrelated." Because each link was evaluated by two different participants, we present the scores per group ("System Links," "Editor Links") and per participant ("participant A," "participant B"). Columns 3 and 6 present the sum of counts for both participants. The chi-square goodness-of-fit test was applied and revealed a statistically significant variation in the observed distribution across all types: (a) χ^2 (4, $N = 164$) = 99.354, $p < .0001$; (b) χ^2 (4, $N = 164$) = 69.293, $p < .0001$; (c) χ^2 (4, $N = 164$) = 165.634, $p < .0001$; and (d) χ^2 (4, $N = 164$) = 158.134, $p < .0001$. We, therefore, reject the null hypothesis that the counts are uniformly distributed across the categories.

We also applied the chi-square test of association to examine whether there is an association between the

TABLE 1. Observed frequencies across the five categories of associated articles.

	System Links (group A)			Editor Links (group B)		
	P_1	P_2	$P_A \cup P_B$	P_1	P_2	$P_A \cup P_B$
Related to main theme	80	69	149	89	84	173
Related to subtopic	34	39	73	51	56	107
Tangentially related	21	25	46	15	19	34
Unrelated	25	27	52	8	2	10
Other	4	4	8	1	3	4

DOI: 10.1002/asi

TABLE 2. Descriptive statistics for editors and across all system configurations.

	Location		Relatedness		Newsworthiness		Interesting Insight	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
System@0.0	2.99	0.87	3.04	0.99	3.31	0.77	2.74	1.01
System@0.1	2.97	0.89	3.05	1.00	3.31	0.80	2.76	1.05
System@0.2	3.10	0.90	3.20	0.99	3.35	0.80	2.86	1.09
System@0.3	**3.18**	0.93	**3.45**	0.96	**3.47**	0.79	**2.92**	1.11
Editors	3.57	0.70	3.80	0.72	3.52	0.70	3.18	0.79

Note. Boldface represents highest performance observed across different system configurations.

participant's group and the type of link. Participants from group B were significantly more likely to find the associated articles related to the main theme (52.7%) or a subtopic (32.6%) of the main article, compared with participants from group A (main theme: 45.4%, subtopic: 22.3%). However, the participants from group A were more likely to perceive the articles as tangentially related (14.0%), compared with participants from group B (10.4%). The system's performance was found to be comparable with that of the editors', with only 15.85% of the embedded links having been reported as unrelated. This is a very encouraging finding, suggesting that our system is scalable and efficient in curating the embedded links.

Table 2 shows the mean and standard deviation scores for participants' assessments of the system-embedded and manually curated links. LEPA's performance is presented across all four precision settings in rows 1 to 4. Four aspects of the links are examined here: (a) whether the anchored text was a good location for the link in the article, (b) whether the anchored text was related to the associated article (the one that the link points to), (c) whether the associated article was newsworthy, and (d) whether the associated article provided interesting insight with respect to the main article. As Table 2 indicates, both types of links received average to good scores, with variations being more evident in terms of location and relatedness. The editors performed better compared with the versions of the system with lower precision, although the performance of the systems with higher precision was clearly comparable. The *U* Mann–Whitney independent groups test also supports this finding for the differences observed between System@0.0 and the editors. A direct comparison between the remaining systems (System@0.1, System@0.2, System@0.3) and the editors was not possible, because the former incorporated only a subset of the links embedded by System@0.0, thus containing an uneven number of links, compared with the number of links that the editors embedded.

The Mann–Whitney test revealed that the manually curated links received statistically significant higher scores than the system links, in terms of location ($U = 8213.5$, $p = .000$, $r = -.35$), relatedness ($U = 7449$, $p = .000$, $r = -.39$), newsworthiness ($U = 11498.5$, $p = 0.02$, $r = -.12$), and interesting insights ($U = 9998$, $p = .000$, $r = -.22$). However, in all cases, the observed differences represent a small effect that accounts for less than 10% of the total variance in our sample. In addition, the more we increased the precision threshold, the more LEPA's performance approximated that of the editors.

The Spearman rank correlation coefficient test was also applied to measure the association between the scores reported for location, relatedness, newsworthiness, and interesting insights, using multiple pairwise comparisons (Table 3). In the majority of cases, Spearman rho revealed a statistically significant positive relationship, indicating an interdependency between the above qualitative aspects of the embedded links. However, the effect size of these relationships was medium only for the cases of the system links, whereas for the editor links it revealed smaller effect sizes.

System performance in terms of standard metrics. Looking at the performance of the system in terms of precision, recall, and *F*-measure in Table 4, we notice that average precision escalates as the threshold value is increased, although this increase has an adverse effect on recall. Apparently, there is a trade-off between introducing fewer but more related links and receiving a larger number of links of heterogenous relevance. In terms of performance, the third column shows the *F*-measure scores, which indicate System@0.0 as the optimum approach. In the next study, we examine which of the two factors (precision, recall) is more strongly correlated with a positive reading experience.

Qualitative analysis. We also performed an inductive, thematic coding of the open-ended questions as part of our qualitative analysis. The coding involved the creation of a classification system based on emerging and meaningful themes. The data were then grouped under these themes, which allowed us to treat them as "of the same type." The analysis was done for the following questions: (a) "[MSTU] Please indicate if the associated article is related to the overall theme of the main article, related to a subtopic within the main article, tangentially related, or unrelated"; (b) "[GPL] Was the underlined text a good place for the link?" (c) "[RTL] Was the underlined text related to the linked article?" and (d) "[INIS] Please state in what ways (if any) the associated article provided interesting insight with respect to the main article?" After the classification system

DOI: 10.1002/asi

TABLE 3. Summary of intercorrelations across all system configurations.

	Location	Relatedness	Newsworthiness	Interesting Insights
System@0.0				
Location	1.000	0.478*	0.373*	0.461*
Relatedness	–	1.000	0.461*	0.570*
Newsworthiness	–	–	1.000	0.550*
Interesting insights	–	–	–	1.000
System@0.1				
Location	1.000	0.505*	0.393*	0.467*
Relatedness	–	1.000	0.472*	0.574*
Newsworthiness	–	–	1.000	0.578*
Interesting insights	–	–	–	1.000
System@0.2				
Location	1.000	0.445*	0.433*	0.461*
Relatedness	–	1.000	0.518*	0.585*
Newsworthiness	–	–	1.000	0.631*
Interesting insights	–	–	–	1.000
System@0.3				
Location	1.000	0.538*	0.487*	0.535*
Relatedness	–	1.000	0.603*	0.610*
Newsworthiness	–	–	1.000	0.717*
Interesting insights	–	–	–	1.000
Editors				
Location	1.000	0.252*	0.254*	0.330*
Relatedness	–	1.000	0.321*	0.372*
Newsworthiness	–	–	1.000	0.509*
Interesting insights	–	–	–	1.000

Note. *Correlation is significant at the .01 level (two-tailed).

TABLE 4. Performance of system across all precision configurations.

	Average Precision	Average Recall	Average *F*-measure
System@0.0	0.5468	**1**	**0.306**
System@0.1	0.5562	0.9467	0.3041
System@0.2	0.5892	0.4719	0.2612
System@0.3	**0.692**	0.2303	0.2456

Note. Boldface represents highest scores.

was concluded, the researcher allocated codes to each of the categories using multiple codes per response. Finally, the responses were blocked per precision level (0.0, 0.1, 0.2, 0.3) to reflect participants' evaluations for each system configuration. The qualitative results reported in this section are referred to by question and number. For example, GPL12 refers to participant 12 responses to the question "[GPL] Was the underlined text a good place for the link?"

Thematic analysis of the responses to "why" the associated article was perceived as related to the overall theme of the main article, to a subtopic within the main article, tangentially related, or unrelated indicated that participants associated higher precision settings with more related links and less redundant or unrelated content. For instance, one participant from System@0.3 noted, "It perfectly hits the same topic, it gives more information about it" (MSTU229), whereas another participant who read a link delivered by System@0.0 reported, "I guess they are connected because they both deal with safety, but one is in the context of natural disaster and one in the context of transportation. They both deal with Florida, also" (MSTU047). Precision settings between 0.0 and 0.2 embedded the most informative links, with precision setting at 0.3 showing some decline most likely because of the strict filtering criteria imposed and the loss of potentially newsworthy information.

A careful examination of responses to "why" the underlined text was a good place for the link revealed a positive relationship between the precision configuration and the suitability of the location, which is in agreement with the quantitative results. A participant from System@0.3 wrote, "I believe it was a good place for the link because the associated article was completely relevant to the underlined text about the health care plan" (GPL182). Participants who evaluated links by systems with high precision configurations also perceived the associated article as being well within the scope of the main article, although some redundant information was occasionally encountered. A participant from System@0.2 noted, "It was okay enough, but I thought the link would lead me to a McQueary only article and not be the same material all over again" (GPL146). Similarly, responses to "why" was the underlined text related to the linked article revealed that high precision settings resulted in topically relevant, informative links, which is also indicated by the findings in Table 2.

Finally, an empirical analysis of participants' responses to the "ways the associated article provided interesting insight with respect to the main article" indicated a positive

DOI: 10.1002/asi

relationship between interesting insight and informative content, that is, news content that provided additional information, insights, and views, as well as support in further understanding the context addressed in the main article. Many of the comments reflected this association, which became more evident as the precision setting increased. For example, a participant from the System@0.3 group of links reported, "It supported the subtopic, that was supporting the main topic being discussed. The news item aims to prove the common notion that people have as wrong. The associated article provides the figures relevant to the same" (INIS026), whereas another noted, "The associated article confirms the news of killing of Gaddafi which was being speculated in the main article. So the associated article is surely providing more insight into the main article" (INIS067).

Assessing the Reading Experience

This experiment aimed at understanding the effects of embedded links from a user's perspective, focusing on the associated reading experience. We used a subset of the data set (presented in Data Set section) and two of the four available system configurations. The motivation was to evaluate the performance of two opposite "extremes" of the system: System@0.0, which applied no precision threshold on the retrieved results; and System@0.2, with a fairly strict but not exaggerated precision threshold. Our choice was informed by the metrics of precision, recall, and *F*-measure, as well as the number of links each system embedded.

Design. We used a between-groups design ("group A," "group B," "group C"), with three independent variables: type of link (two levels: "system-embedded," "manually curated"), publication date (four levels: "19/10/2011," "20/10/2011," "16/11/2011," "17/11/2011"), and precision configuration (two levels: 0.0, 0.2). The levels of the independent variables were manipulated in the manner described in the Design section.

Each participant took part in one condition (one article-link combination) and assessed four different aspects of the task: (a) the main article, (b) the associated articles, (c) the links, and (d) the reading experience. The participants in groups A and B evaluated the system-embedded links with precision configurations 0.0 and 0.2, respectively, whereas the participants of group C evaluated the manually curated links. The dependent variables related to the main article were: (a) informativeness, (b) newsworthiness, and (c) similarity to other news read online. The dependent variables examined in terms of the associated articles were: (a) relatedness to the main article, (b) newsworthiness, (c) interesting insight with respect to the main article, and (d) topical coverage. Regarding the embedded links, the variables examined involved the adequacy of links and the number of links clicked in a real-life scenario. Finally, the reading experience was evaluated based on a comparison with the average online news reading experience.

TABLE 5. Titles of the news articles used in the reading tasks.

19/10/2011: Turkey launches incursion into Iraq.
20/10/2011: Gadhafi's son Seif al-Islam captured and wounded.
16/11/2011: Zooming in on the Olympus scandal.
17/11/2011: Police: Penn State asst. didn't tell us of abuse.

Tasks. We selected four articles, each from a different date, while considering the number of embedded links it contained and its mean precision score. For each of these articles there was one version prepared by the editors and two by the system, using the precision thresholds 0.0 and 0.2. The number of embedded links varied between 2 and 11 per precision configuration and per date of collection. In all cases, the links that appeared in the System@0.2 condition were a subset of the links embedded by the System@0.0. Table 5 shows the dates and the titles of the articles that were used.

Each article was read by 10 different participants. The participants were instructed to read the news content, as well as the content of all the embedded links found in the text. In this experiment, we presented the articles using rich HTML format, because our goal was to capture the online news reading experience in a realistic environment. The content of each link was shown in a pop-up box allowing for a quick evaluation. Upon completing the task, the participants were redirected to an online questionnaire. Similarly to the previous experiment, we took measures to discourage cross-site scripting attacks or automatic and malicious completion of the questionnaires.

Questionnaire. A post-task questionnaire was used to elicit information on several aspects of the task, such as the main and associated articles, the quality of the embedded links, and the participants' reading experience (see Questionnaire section). In addition, we introduced follow-up questions, conditional upon each response, to elicit the reasons behind the participants' choices.

Procedure. The participants were asked to complete the task in a single sitting, including the questionnaire. They were also informed of their option to opt out from the task at any point, without being compensated. The duration of the tasks lasted between 20 and 75 minutes, and the payment for participation was between $1.60 and $5.50, depending on the number of embedded links each article contained.

Participants. A total of 120 participants were recruited, who spent an average of 38 minutes on each task and provided a total of 68 hours of labor. They were randomly distributed into three groups of 40 people. They were of mixed ethnicity and educational background, and were all proficient with the English language. Group A consisted of 50% male and 50% female participants mainly younger than 41 years (87.5%), with the largest group between the ages of 24 and 29 (27.5%). The majority (97.5%) indicated reading

DOI: 10.1002/asi

TABLE 6. Descriptive statistics for editors and across all system configurations.

	Relatedness		Newsworthiness		Interesting Insights		No. of Links		Topical Coverage	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
System@0.0	3.85	0.77	**3.60**	0.87	**3.77**	0.94	2.65	0.83	3.97	0.80
System@0.2	3.50	0.90	3.55	1.08	3.37	1.10	**2.70**	0.75	3.62	1.10
Editors	**3.90**	0.74	3.55	0.98	3.55	1.13	2.65	0.94	**4.00**	0.84

Note. Boldface indicates highest performance observed across different system configurations.

the news online between a few times per week and several times a day. Group B consisted of 52.5% male and 47.5% female participants mainly younger than 53 years (97.5%), with the largest group between the ages of 24 and 29 (30%). The majority of the participants (55%) also indicated reading the news online several times a day. Finally, group C consisted of 60% male and 40% female participants mainly younger than 35 years (85%), with the largest group between the ages of 24 and 29 (40%). The majority (92.5%) indicated reading the news online several times a day.

The Kruskal-Wallis test was applied to compare the three groups for differences in relation to demographics (age, gender, educational level, proficiency with English, frequency of reading the news). To take an appropriate control of Type I errors, we applied a Bonferroni correction, and thus all effects are reported at .016 and .025 levels of significance accordingly. The only statistically significant difference found was with respect to the level of education [$H(2) = 8.847$, $p = .012$]. The post hoc tests revealed that the differences between groups A and C ($U = 530.5$, $p = .006$, $r = -.3$) and groups B and C ($U = 567$, $p = .019$, $r = -.26$) were statistically significant, although for both the effect size was very small.

Results. We evaluate the performance of our automated hyperlinking approach against the editors (72 links: 33 from System@0.0, 16 from System@0.2, and 23 manually curated). The performance was measured using precision, recall, and the *F*-measure, and as ground truth we used the participants' scores. Questions that ask for user rating on a scale of 1 to 5 represent stronger perception with high scores and weaker perception with low scores. The Kruskal-Wallis test was used to establish the statistical significance ($p < .05$) of the differences observed in the experimental findings. In addition, we perform a correlation analysis and examine the nature of the relationship (positive, negative, or no linear relationship) among the dependent variables that characterize the quality of the links, across all groups of articles. Although our analysis does not establish cause-and-effect relationships, it indicates how and to what degree the examined variables are associated with each other, thus providing further insights into the extent the system and the manually curated links satisfied the qualitative criteria.

Quantitative analysis of main and associated articles. To evaluate the main article across all groups, the participants were asked to answer the following questions: (a) "Did you find the article informative?" (b) "How interesting did you find the article you read?" and (c) "Was the article you read similar to other news you usually read?" With respect to the first question, the participants' scores revealed that the main article was perceived as very informative ($M = 4$, $SD = 0.0$). For the second question, the participants reported the article to be somewhat to very interesting ($M = 3.48$, $SD = 1.07$); whereas in response to the third question, they claimed that it was somewhat similar to other news they usually read ($M = 3.05$, $SD = 1.13$). The results indicate that the news articles we used as our data set were a fair approximation of what users read online on any given day. Also, the scores that were assigned to *interest* and *informativeness* indicate that the latter variables were not adversely affected by our experimental setup.

Table 6 shows the mean and standard deviation scores for participants' assessments of the system-embedded links and the manually curated ones by professional editors. Five aspects of the links are examined: (a) the relatedness of the anchored text to the associated article, (b) the newsworthiness of the associated article, (c) the degree to which the associated articles provided interesting insights with respect to the main article, (d) the adequacy of the number of embedded links, and (e) the level of topical coverage that the associated articles offered. The Kruskal-Wallis test was applied on all five conditions, but it did not indicate a statistically significant difference. With the exception of *relatedness* and *topical coverage*, the system links received higher scores compared with those of the editors. System@0.0 held the best performance in terms of embedding newsworthy links and articles that provided interesting insights.

A correlation analysis was also applied to examine the association between relatedness, newsworthiness, interesting insights, number of links, and topical coverage. The Spearman rank correlation coefficient test revealed several statistically significant positive relationships (Table 7), which were more evident (larger effect size) for System@0.2, compared with System@0.0 or the editors. This suggests that the links embedded by System@0.2 satisfied in parallel and to a greater extent the earlier qualitative criteria, when embedded successfully. On the other hand, links curated by System@0.0 or the editors, even when accurate, did not achieve the same effect.

DOI: 10.1002/asi

TABLE 7. Summary of intercorrelations across all system configurations.

	Relatedness	Newsworthiness	Interesting insights	No. of links	Topical coverage
System@0.0					
Relatedness	1.000	0.396*	0.120	–0.149	0.159
Newsworthiness	–	1.000	0.558**	–0.152	0.510**
Interesing insights	–	–	1.000	–0.253	0.355*
No. of links	–	–	–	1.000	0.012
Topical coverage	–	–	–	–	1.000
System@0.2					
Relatedness	1.000	0.665**	0.536**	–0.060	0.508**
Newsworthiness	–	1.000	0.554**	0.083	0.561**
Interesing insights	–	–	1.000	0.368*	0.692**
No. of links	–	–	–	1.000	0.340*
Topical coverage	–	–	–	–	1.000
Editors					
Relatedness	1.000	0.561**	0.406**	0.035	0.234
Newsworthiness	–	1.000	0.663**	0.062	0.353**
Interesing insights	–	–	1.000	0.094	0.346*
No. of links	–	–	–	1.000	0.256
Topical coverage	–	–	–	–	1.000

Note. *Correlation is significant at the .05 level (two-tailed). **Correlation is significant at the .01 level (two-tailed).

Qualitative assessment and emerging themes. A qualitative analysis of the open-ended questions was also performed through a process of inductive, thematic coding. The coding process involved the creation of a classification system that imposed a particular order on the data for four different questions: (a) "[NOL] In general, you found the number of associated articles to be . . ."; (b) "[TC] Overall, did you feel that the associated articles provided a good topical coverage for the main article?" (c) "[COE] Compared to your average experience reading online news, how would you rate this new reading experience?" and (d) "[IMA] Please state in what ways (if any) the associated articles provided interesting insights with respect to the main article." The analysis was concluded by allocating codes to each of the categories, using multiple codes per response. The qualitative results reported in this section are referred to by question and number. For example, TC77 refers to participant's 77 responses to the question: "[TC] Overall, did you feel that the associated articles provided a good topical coverage for the main article?"

An independent researcher was asked to repeat the process following the same classification system to determine how well the implementation of the coding scheme worked. An inter-rater reliability analysis using the Kappa statistic was applied to determine the level of consistency between the ratings of two researchers. The results of the inter-rater analysis were found to be $Kappa = 0.9$ with $p < .000$ for NOL, $Kappa = 0.887$ with $p < .000$ for TC, $Kappa = 0.931$ with $p < .000$ for COE, and $Kappa = 0.973$ with $p < .000$ for IMA. The Kappa scores were all found to be statistically significant and suggest an almost perfect agreement beyond chance. Following this initial step, some of the code definitions were revised and clarified to improve coding consistency, and researchers reached agreement on those responses where their coding conflicted.

Thematic analysis of responses to "why" the associated articles provided interesting insights revealed that participants associated interesting insights with good topical coverage and informativeness, as well as the degree that an article offered a broader perspective. One participant from group A noted: "The associated articles helped to put the main article in context by providing background on the events that led up to the event described in the main article as well as some reaction from key political leaders to the events that were unfolding" (IMA8), whereas another participant from group C reported: "The associated articles were insightful, as some provided elaborate details and others provided different perspective of the topic covered in the main article" (IMA101).

In terms of the number of embedded links, System@0.2 achieved the best performance. An examination of the open-ended responses indicates dissimilar patterns. Participants from groups A and C appeared more likely to be affected by articles with excessively long and redundant content, as well as lack of relatedness. One participant wrote: "They were too many, being mostly quite long, in some cases more than half the length of the main article, and sometimes they repeated the same identical information" (NOL80). On the contrary, the responses of participants from group B reveal that the number of links was adequate and the associated content they pointed to was perceived as less redundant. NOL61 wrote, "It gave just enough break up from the main article to allow me to reflect on it while receiving new information." Lastly, the editors were more successful in curating links that provided good topical coverage and had a substantial thematic connection. Responses to the "why" portion of this question suggest

DOI: 10.1002/asi

that this effect was achieved by introducing articles with less redundant or unrelated content. One participant noted, "The articles were on topic, and helped put the original news story in perspective" (TC77).

Online news reading experience. To understand better the effects of embedded links from a user's perspective and the associated reading experience, we also inquired about the following two aspects of the task: (a) "Compared to your average experience reading online news, how would you rate this new reading experience?" and (b) "How many of the links that appear in the text would you have clicked to read, if it was not obligatory to open them all as part of this study?" For the first question, the scores of the manually curated links were marginally higher ($M = 3.35$, $SD = 1.001$) than those of System@0.0 ($M = 3.325$, $SD = 0.971$) and System@0.2 ($M = 3.225$, $SD = 0.8002$). On average, the reading task was perceived as the same, or somewhat better, compared with their average online news reading experience, across all groups. Among the most prominent themes that emerged from the analysis of the open-ended responses was topical coverage, link presentation, and content volume. The responses indicate that these three factors were balanced, which resulted in a positive news reading experience. One participant in group B wrote: "In general, online news doesn't cover the topic entirely. Even the related news is not perfectly related. So, this was somewhat better than them" (COE16). In the second question, the participants' scores favored the editors ($M = 2.9$, $SD = 1.2567$) over System@0.0 ($M = 2.6$, $SD = 1.104$) and System@0.2 ($M = 2.475$, $SD = 1.3772$), indicating that they would have clicked on most of the embedded links if it was not a mandatory condition of the task. However, the Kruskal-Wallis test did not reveal any statistically significant difference for any of the above conditions.

Finally, an empirical analysis of participants' "additional comments" section made evident several important themes, with the most predominant being "relatedness," compared with "large volume" and "redundancy." Many of the comments in our study reflect the assumption that the associated articles should be directly related to the main article, with little tolerance for semi-related or unrelated information. In addition, several participants indicated a preference for well-written, informative articles that contained less text, and ignored the extremely long ones. A high number of embedded links was also considered as a distraction, and the participants were more reluctant to read them all. This finding establishes the importance of precision over recall.

Overall, our experimental findings indicate that the system's performance was found comparable with that of the editors. Also, the quantitative results were in agreement with the qualitative ones in terms of the system and editor comparison. This is an encouraging finding, considering that the main contribution of our automated approach is scalability. News editing is a challenging task for humans, and support from an automated system could improve the efficiency and performance of the editors.

Discussion

We conducted two large and thorough experiments via MTurk to evaluate our system-embedded links against links manually curated by professional editors. The two experiments were designed on the basis that 64.85% of the system-generated links were assessed as "good enough," providing evidence of LEPA's potential to support the link curation process, but only 7% of the generated links were common to LEPA and professional news editors, indicating that the system links were complementary to those of the professional editors. Our two experiments, therefore, evaluate the system-generated links both in terms of their quality as compared with those generated by professional editors *and* with respect to their overall effect on the news reading experience. This section discusses our main findings.

The first experiment allowed us to perform a close examination of our automated approach to link-generation using standard quantitative metrics, and qualitative judgments from human evaluators, for each link individually. Our evaluation reveals that the editors had an average to good performance, whereas our system had an overall average performance. When we treat the manual links as a gold standard, the results are encouraging, indicating that LEPA is comparable with that across several facets of the news reading experience (e.g., relatedness of the anchored text with the associated article, newsworthiness, offering interesting insights). Our correlation analysis also indicated a positive association between several of the qualitative aspects, for example, *location*, *relatedness*, *newsworthiness*, and *interesting insights*. The effect size of these relationships was medium only for the cases of the system links (especially for those cases with a stricter precision configuration), whereas for the editor links, the analysis revealed smaller effect sizes. When examining the system performance using standard metrics, we observe that the *F*-measure scores decline over high precision values. Apparently, assigning a high value to the precision threshold acts as a trade-off for recall and vice versa.

Furthermore, analysis of the open-ended responses revealed several interesting findings. For instance, the participants associated higher precision settings with more related links and less redundant or unrelated content. The precision configuration was also found to be positively associated with the suitability of the hypertext's location, as well as the topical relevance of the associated article, which is in agreement with our quantitative results. In addition, many of the participants' comments reflected a relationship between interesting insight and informative content, which became more evident as the precision setting increased. Precision settings between 0.0 and 0.2 embedded the most informative links, with precision setting at 0.3 showing some decline most likely because of the strict filtering criteria imposed and the loss of potentially newsworthy information. This finding confirms the importance of topically relevant content that highlights different facets of the news and stimulates the readers to consider divergent interpretations.

DOI: 10.1002/asi

The second experiment aimed at understanding the effects of embedded links from a user's perspective, focusing on the associated reading experience. Guided by the findings from the previous study, we selected and evaluated a further two opposite "extremes" of our system: a version that applied no precision threshold on the retrieved results and a version with a fairly strict, but not exaggerated, precision threshold. The experimental results suggest that LEPA, despite its average performance, proved to have the potential to assist the editors in the link-curation process (and, in some cases, even perform better), such as delivering newsworthy links or providing interesting insights. The correlation analysis we ran also indicated the existence of positive interdependences among several qualitative aspects of the links, like *relatedness*, *newsworthiness*, *interesting insights*, *topical coverage*, and other. Similar to the previous study, the correlation test revealed that the system with the highest precision value was able to satisfy in parallel and to a greater extent the earlier qualitative criteria, compared with the system with no precision threshold or the editors, which did not achieve the same effect.

Several themes emerged from the qualitative analysis of the participants' responses. We identified a correlation between the theme *interesting insights* and *good topical relevance*, *informativeness*, and *broader perspective*. It appears that an article with the earlier qualities is more likely to offer interesting insights with respect to the main article. Another association was established between the *number of links* and *redundancy*, as well as *text volume* and *relatedness*. Having articles that provide related yet unseen content with reasonable length can affect the way news readers perceive the availability of embedded links. Moreover, *topical coverage* was found to be associated with *redundancy* and *relatedness*, which makes the reasonable case that unrelated and redundant content does not afford substantial topical coverage. Finally, the themes of *topical coverage*, *link presentation*, and *content volume* are some of the main factors that contribute to a *positive news reading experience*. This qualitative finding is of significance, particularly to professional news editors, because it reflects the qualities of online news with which readers are less willing to compromise.

Finally, although the use of Amazon MTurk as a platform for online recruitment took much of the environmental control away that we could have maintained in a laboratory setting, it also allowed for more rapid testing and a larger and more diverse participant base than the average user study. This was particularly important for showcasing the benefits of LEPA, considering its relative simplicity. Particular attention was paid to the experimental method. We first had to take preventive measures to discount low-quality responses and undesirable participants that is known to arise with crowd-sourcing platforms. Carefully designed questionnaires allowed us to elicit findings that not only provided information on the quality of the automatically generated links, but also on the overall news reading experience. In addition, the MTurk workers made excellent participants, because their responses to open-ended questions were thoughtful and contributed significantly to the interpretation of our results. The rigorous evaluation method used in this article and our results go beyond LEPA; both are relevant to any website designer interested in promoting higher engagement through click-through experience.

Conclusions

The manual creation of hyperlinks is a time-demanding and challenging task, especially for large online news providers where editors are expected to process a significant number of news articles on a daily basis. Consequently, any effort toward automating the link curation process could go a long way to improve their efficiency and performance. In this article, we presented LEPA, a fully automated approach to detecting events in news articles and linking them to relevant past articles. Unlike other automated systems that focus on link detection and disambiguation, LEPA does not require any training data or human intervention, nor is it limited to a specific resource (e.g., Wikipedia), making it a generalizable and attractive solution for many domains.

We performed a rigorous evaluation of LEPA to assess the automatically embedded links against links manually curated by professional editors. Three independent evaluations carried out provided several quantitative and qualitative findings that helped understand better the curation of links from the perspective of users and professional news editors, and offered empirical support to the *why* and *how* of the news reading experiences. Our findings indicated that the performance of our automated linking approach is comparable with that of the editors, across all qualitative aspects of the link curation process we examined. For our setting, high-precision configurations facilitated better news reading experiences, unlike lower thresholds that resulted in a larger number of links and provided access to a plethora of information. In other words, *less is more*. Our correlation analysis further supports this conclusion, and also highlights the effect of high precision on the extent and depth to which the system links meet the qualitative criteria, something not observable on the same degree for the editors or other system configurations. In a real use-case scenario, the final decision might ultimately be for the editors to make, but our experimental findings indicate that the proposed system, with its massive scalability being its greatest asset, can support the process of identifying interesting links and fulfill its purpose in reducing manual effort.

In conclusion, the proposed automatic approach to link curation offers a news-reading experience that is comparable with that of manually curated links by professional editors, with the potential to reduce the manual labor required for the task. A number of directions are left to be explored in future work. These include testing the linking capabilities of LEPA on other resources, such as medical reports, encyclopedic articles, or Wikipedia documents. Finally, our experimental method can be expanded to account for online behavior

DOI: 10.1002/asi

metrics and engagement measures like cursor behavior. A potential future application, with the help of such metrics, would be to determine in real-time the sections of a news article that are of particular interest to the reader and perform a dynamic, on-the-fly adaptation of the embedded links, thus promoting more engaging news-reading experiences.

References

Agarwal, D., Chen, B.-C., & Elango, P. (2009). Explore/exploit schemes for web content optimization. In Proceedings of the 2009 Ninth IEEE International Conference on Data Mining (ICDM '09) (pp. 1–10). Washington, DC, USA: IEEE Computer Society.

Allan, J. (1995). Automatic hypertext construction (PhD thesis). Cornell University.

Attfield, S., Kazai, G., Lalmas, M., & Piwowarski, B. (2011). Towards a science of user engagement (Position Paper). In *WSDM Workshop on User Modelling for Web Applications*.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

Bluestein, W. (1999). Hypertext versions of journal articles: Computer aided linking and realistic human evaluation (PhD thesis). University of Western Ontario.

Bron, M., Huurnink, B., & de Rijke, M. (2011). Linking archives using document enrichment and term selection. In Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries (TPDL'11) (pp. 360–371). Berlin, Heidelberg: Springer-Verlag.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 708–716). Prague, Czech Republic: Association for Computational Linguistics.

Demartini, G., Difallah, D.E., & Cudré-Mauroux, P. (2012). Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In Proceedings of the 21st International Conference on World Wide Web (WWW '12) (pp. 469–478). New York, NY, USA: ACM.

Ellis, D., Furner, J., & Willett, P. (1996). On the creation of hypertext links in full-text documents: Measurement of retrieval effectiveness. JASIS, 47(4), 287–300.

Fissaha Adafre, S., & de Rijke, M. (2005). Discovering missing links in Wikipedia. In Proceedings of the 3rd international workshop on Link discovery (LinkKDD '05) (pp. 90–97). New York, NY, USA: ACM.

Gardner, J.J., & Xiong, L. (2009). Automatic link detection: a sequence labeling approach. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2–6, 2009 (pp. 1701–1704).

Green, S.J. (1997). Building newspaper links in newspaper articles using semantic similarity. In Proceedings of the Natural Language Processing and Information Systems, 3rd International Conference on Applications of Natural Language to Information Systems, NLDB 1997, Vancouver, British Columbia (pp. 178–190).

Green, S.J. (1998). Automated link generation: Can we do better than termrepetition? Computer Networks, 30(1–7), 75–84.

Hakan, C., Arapakis, I., Donmez, P., & Lalmas, M. (2012). Automatically embedding newsworthy links to articles. CIKM, 1502–1506.

He, J., & de Rijke, M. (2010). A ranking approach to target detection for automatic link generation. In Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010) (pp. 831–832). Geneva, Switzerland.

He, J., de Rijke, M., Sevenster, M., van Ommering, R. C., & Qian, Y. (2011). Generating links to background knowledge: a case study using narrative radiology reports. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11) (pp. 1867–1876). New York, NY, USA: ACM,

Huang, W.C., Geva, S., & Trotman, A. (2009). Overview of the inex 2009 link the wiki track. In Proceedings of the Focused Retrieval and Evaluation, 8th International Conference on Initiative for the Evaluation of XML Retrieval (INEX'09) (pp. 312–323). Berlin, Heidelberg: Springer-Verlag.

Jijkoun, V., Khalid, M.A., Marx, M., & de Rijke, M. (2008). Named entity normalization in user generated content. In Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data (AND '08) (pp. 23–30). New York, NY, USA: ACM.

Kittur, A., Suh, B., Pendleton, B.A., & Chi, E.H. (2007). He says, she says: conflict and coordination in wikipedia. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07) (pp. 453–462). New York, NY, USA: ACM.

Knoth, P., Zilka, L., & Zdrahal, Z. (2011). Using explicit semantic analysis for cross- lingual link discovery. In Proceedings of the International Workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA) at The 5th International Joint Conference on Natural Language Processing (IJC-NLP 2011).

Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., & Chang, Y. (2011). Learning to model relatedness for news recommendation. In Proceedings of the 20th International World Wide Web Conference (WWW '11) (pp. 57–66). New York, NY, USA: ACM.

Mason, W., & Suri, S. (2010). Conducting behavioral research on amazon's mechanical turk. Behavior Research Methods, 1–23. doi: 10.3758/s13428-011-0124-6.

Mihalcea, R., & Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6–10, 2007 (pp. 233–242).

Milne, D.N., & Witten, I.H. (2008). Learning to link with wikipedia. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26–30, 2008 (pp. 509–518).

Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., & Olmstead, K. (2010). Understanding the Participatory News Consumer. Pew Internet & American Life Project.

Shen, W., Wang, J., Luo, P., & Wang, M. (2012). Linden: linking named entities with knowledge base via semantic knowledge. In Proceedings of the 21st International World Wide Web Conference (WWW '12) (pp. 449–458). New York, NY, USA: ACM.

Suchanek, F., Kasneci, G., & Weikum, G. (2012). Yago: A large ontology from Wikipedia and WordNet. Journal of Web Semantics, 6(3), 203–217.

Wang, Z., Li, J., Wang, Z., & Tang, J. Cross-lingual knowledge linking across wiki knowledge bases. In Proceedings of the 21st International World Wide Web Conference (WWW '12) (pp. 459–468). New York, NY, USA: ACM.

DOI: 10.1002/asi