VisualWikiCurator: Human and Machine Intelligence for Organizing Wiki Content

Nicholas Kong*, Ben Hanrahan*, Thiébaud Weksteen†, Gregorio Convertino*, Ed H. Chi*
Palo Alto Research Center (PARC)*, Xerox Research Centre Europe†
Palo Alto, CA 94304, USA | 38240 Meylan, France
{nkong, bhanrahan, convertino, echi}@parc.com, thiebaud.weksteen@xrce.xerox.com

ABSTRACT
Corporate wikis are affected by poor adoption rates. The high interaction costs required to organize and maintain information in these wikis are a key factor that limits broader adoption. We present VisualWikiCurator, a wiki extension designed to lower such costs by (a) recommending new content to easily update a wiki page, and (b) extracting structured data from the wiki page while providing new alternative visualizations of the data. The visualizations of extracted semantic data act both as alternative views and as tools to organize the page content. Since no information extraction algorithm is perfect with generic unstructured data, we use a *mixed-initiative* approach to allow users to refine machine-extracted metadata and easily re-organize the content in wiki pages.

Author Keywords
Corporate wikis, organization, visualization, Web 2.0

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Keywords
Design

INTRODUCTION
Although pervasive in organizations, wikis are affected by poor adoption rates [11]. A key factor in these low adoption rates is the high cost for users to organize and maintain wikis, which often results in data sparsity and poor organization. The cost associated with these actions is mostly due to the static, or non-interactive, nature of wikis and their tendency to become disorganized and out of date [2, 9]. Many of these pages thus remain incomplete, while consuming the resources of and confusing those who visit (e.g., [5]). We propose a system that seeks to reduce the cost of updating and organizing wiki pages by combining human and machine intelligence.

IUI 2011, February 13–16, 2011, Palo Alto, California, USA.

The poor organization and heterogeneity of data on corporate wikis obstructs users in better consuming them. Corporate wiki pages vary widely in type (organized lists, project updates, tutorials), level of curation, and formatting styles. The sparsity and heterogeneity of data make automated techniques, such as entity extraction algorithms, even more error-prone than when these techniques are applied to richly populated public wikis (see [9]). These constraints motivate our mixed-initiative approach.

We address these problems in three ways. First, to help populate the wiki and to keep it updated, our system recommends dynamic content, pulled from email and RSS feeds; this content may contain updates relevant to the current wiki page. We borrow this idea from other tools that presents other related content in context. For example, targeted advertisements (e.g., Google Ads) recommend contextual dynamic content given the content of a "static" webpage (e.g., an e-mail).

Figure 1. Interface overview. (top) The visual bridge, displays alternative views (a timeline here) with metadata extracted from the wiki page and dynamic content. (bottom) Relevant dynamic content (emails here) is recommended at the bottom of each page.

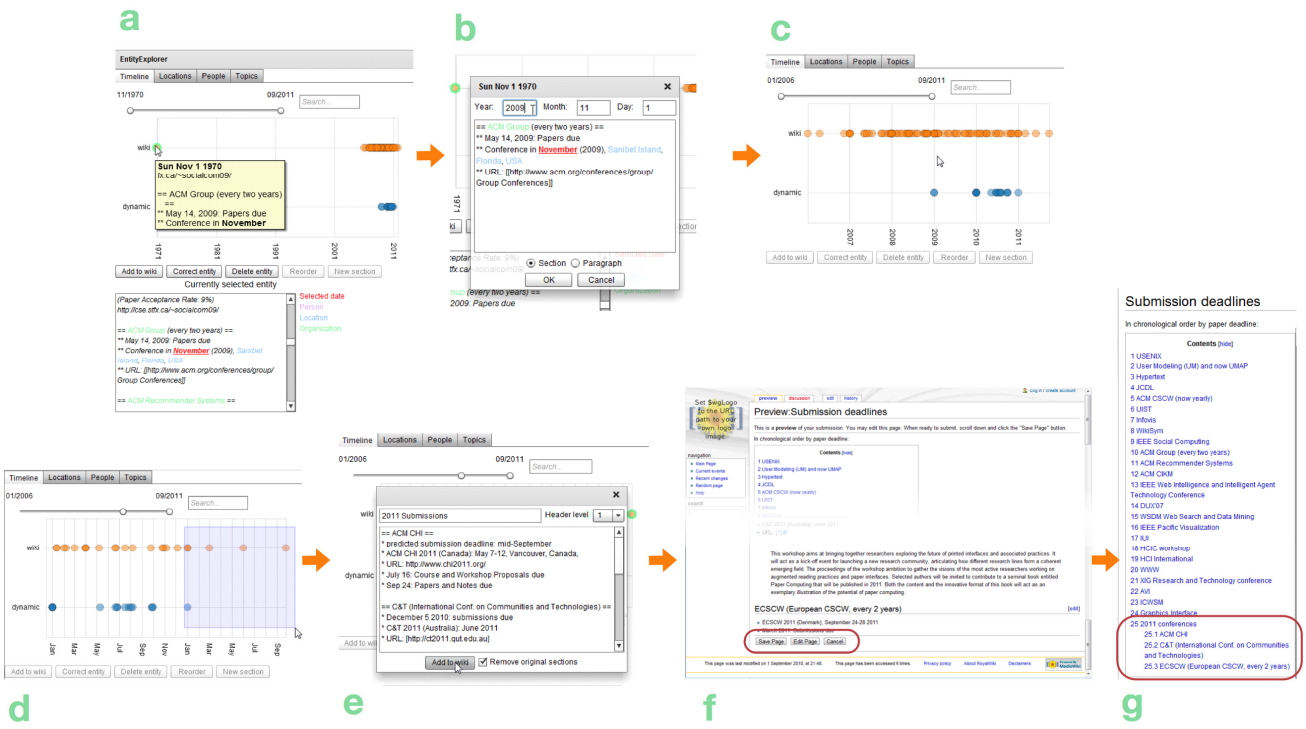

Figure 2. A user organizes a wiki page containing conference deadlines. See the Scenario section for details. (a-c) The user corrects an entity. (d-g) The user selects entities to create a new section on the wiki page.

Second, to assist with the organization of content in wikis, we implement direct manipulation of visualizations. We built on ideas from several tools already aimed at helping users visually organize content in wikis, such as Vispedia [3], and visual data analysis tools, such as Jigsaw [15] and Entity Workspace [1]. Finally, we draw from past research on supporting organization in wikis via natural language processing for semantic web tools, such as the semantic Mediawiki extension (http://semantic-mediawiki.org), Wikulu [8], and Woogle [7]. In particular, SAVVY Wiki [12] helps to organize fragmentary items across wiki pages.

Third, we use mixed-initiative [11] as our general approach to mitigate the inevitable errors from automatic extraction of semantic information from unstructured data. In our system, user actions are combined with the results of automatic functions. This approach has been used in past systems that allow individuals to easily categorize data and refine machine-inferred metadata [13]. But only a few tools support these functionalities in collaborative contexts, such as Hoffman's work on infoboxes in Wikipedia [9].

DESIGN

The system consists of a MediaWiki extension that interfaces with a Django backend that stores, indexes and serves dynamic content. Figure 1 illustrates the two main components of our system. The first component recommends content that is relevant to the current wiki page for future inclusion and organization (Figure 1 bottom). The second component extracts metadata from the content and visualizes this metadata in a "visual bridge" (Figure 1 top). Via the bridge, multiple users can analyze the content from multiple views, correct any erroneous metadata extractions, and organize the wiki page. We extract entities as our metadata. We detail each component separately below.

Recommending dynamic content

In order to help populate the wiki and to keep pages up-to-date, we lower the cost of foraging and collecting new relevant information. Our system identifies and appends relevant external content, such as email and RSS feeds, at the bottom of each wiki page. We term this external content "dynamic" to emphasize that its information is frequently updated, in contrast to the information in the wiki page which is relatively "static". The system may accept different kinds of dynamic sources (emails, feeds, tweets, etc.). As a new email arrives, we associate it with the most relevant wiki page. For feeds, we fetch them on a regular basis and apply the same process.

This backend includes a recommendation module that computes similarities between dynamic and static content, which we use to place relevant external content next to each wiki page. The module uses the BM25 similarity metric [14] between the two texts (considered as bags of words). This metric allows us to learn from user interactions to produce personalized rankings. For example, if a user recently modifies a specific wiki page and is also the author of a new e-mail, then the algorithm moves this e-mail up in

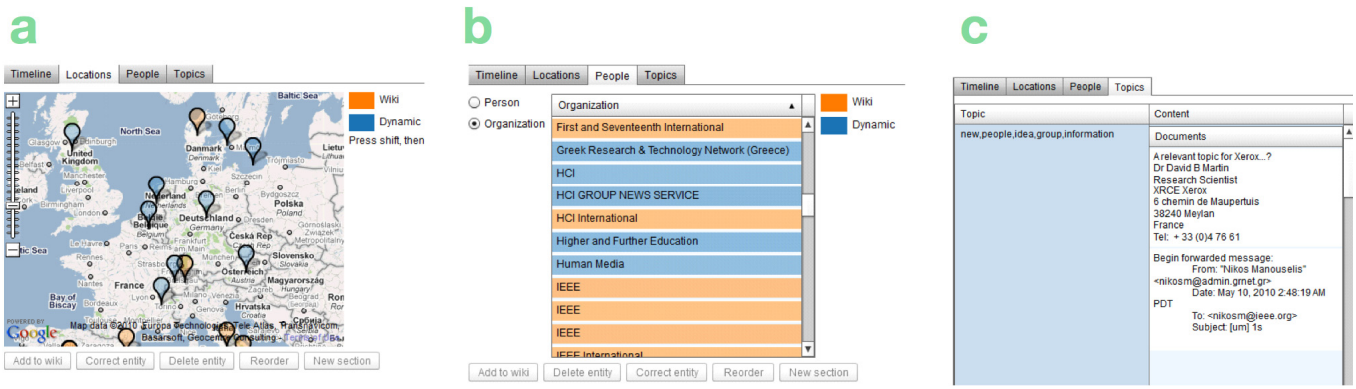

Figure 3. Other views in the visual bridge. The timeline view is shown in Figure 2. (a) Location view. (b) Person/Organization view. (c) Topical view.

the ranking associated with this wiki page. However, we leave this feature to future work.

Visual bridge: metadata extraction and organization

The second component, the visual bridge, automatically extracts metadata (i.e., entities) from the dynamic and static content and provides multiple views of the metadata. We used ANNIE, the entity extractor shipped with GATE [4]. Our system extracts dates, people, organizations, and locations from both the dynamic and static content.

The visual bridge provides a *mixed-initiative method for organization*: the system extracts semantic data and recommends ways to sort the wiki based on user selection. Then the user can preview and correct these suggestions before committing the changes (Figure 2(a-c)). The visual bridge was built in Flash using the flare toolkit (http://flare.prefuse.org).

The metadata are visualized in multiple alternative views: a timeline for dates, a sortable table of people and organizations, a map for locations, and a topical view showing clusters from a topical model (see Figure 3). For each wiki page, the system chooses the view to display first. Currently a simple heuristic is used: the system counts the number of entities extracted for each type and chooses the view with the most entities (e.g., a timeline in Figure 2a).

Finally, the system allows users to reorganize the wiki page *through* the visual bridge. For example, the user can select a set of entities in the timeline and re-sort the wiki page by date based on those entities (Figure 2(d-g)). A similar operation is possible in the map and person/organization view.

Mixed-initiative to extract, correct, and use metadata

In the timeline, person/organization, and map views, the user can review and correct extracted metadata. For example, she may correct "Athens" to "Athens, GA". This feedback could allow the system to learn to correct similar extraction errors, even though the current prototype does not yet do this.

Each entity is also associated with a piece of text, suggested by the system. Entities extracted from the static content are associated with the wiki section that contains them, while entities extracted from the dynamic content are associated with the sentence that contains them. Users can also manually specify the text that is associated with a particular entity (Figure 2b), or switch between system-recommended associations (e.g., a wiki section or paragraph). The views in the visual bridge allow the user to interactively reorder the page text using the entities as "handles", and finally correct and commit the new text (Figure 2(d-g)).

SCENARIO: ORGANIZE CONFERENCE DEADLINES

In our fieldwork with communities of professionals in two companies (a large business enterprise and a research center), we reviewed the pages of about fifteen corporate wikis. The wikis generally included organized lists of items (work tools, people by area, meeting notes or schedules). This scenario represents an example of curation work that typically occurs around a specific exemplar of these pages and serves to illustrate a typical use case for our system.

Sonia is browsing a wiki page that lists conference deadlines relevant to her team. On behalf of her collaborators, she decides to use VisualWikiCurator to group the deadlines that will occur in 2011 into a new section, including any recommended new calls not yet listed in this wiki page. Figure 2 shows a flow chart with her sequence of actions:

- [a] Sonia begins work in the timeline view of the visual bridge. She sees that a date that has been assigned a year of 1970, hovers her mouse over the mark to see its content (see yellow tooltip in Figure 2a), and notices that the date was incorrectly extracted: the year should be 2009.
- [b] She clicks the "Correct Entity" button (Figure 2b), changes the date, and confirms.
- [c] Timeline updates with corrected date (Figure 2c).
- [d] She filters down by using the slider at the top, selects a specific interval of dates in 2011 with her mouse (see the blue selection box in Figure 2d), and clicks the "Add new section" button.
- [e] This opens a dialog box prompting for the new section name and a preview of the section content in

edit mode (Figure 2e), which she quickly revises.

- [f] Next, the system opens a preview window of the wiki page with the changes (Figure 2f). She scrolls down the page, checks the new section temporarily added at the bottom and clicks the "Save Page" button.
- [g] Finally, the system redirects her to the published wiki page with the new section (Figure 2g).

FUTURE WORK

We would like to extend the mixed-initiative capabilities of our system. The system could do searches or extract entities from the public web (e.g., using URLs in the wikitext) to fill in missing information. The system could also learn patterns from examples (e.g., list of CFPs in Figure 2), create tables from schemas, or recommend visualizations to embed. We have also begun exploring algorithms for automatically generating a multi-section page based on a set of dynamic content (such as a project page from a set of e-mails). This work could be further extended to the generation of meta-pages which group similar pages together.

Our system does not currently store metadata or user corrections. We plan on implementing this in the future, as it is a shortcoming that hampers collaboration.

Finally, we plan to evaluate our system, both to ensure that the system is easy to use and reduces the cost of organizing content in wikis. We would like to run lab studies to examine its effect on the time required to organize pages and the quality of those pages. We would also like to run longitudinal studies to examine its effect on increasing the total number of contributions.

CONCLUSION

Several new tools have enabled knowledge workers to quickly forage for large numbers of independent pieces of information, such as search results, RSS feeds and blogs. Other tools have enabled groups or large communities to collaboratively edit fully structured content in a shared space (e.g., wiki pages, Google Docs). In contrast with the abundance of the tools of both classes, very few tools are available to assist workers in the in-between work of filtering, abstracting, and organizing low-level pieces of information into intermediate shared products. In other words, we lack tools that lower the costs for collaborative sensemaking and organization.

Particularly for corporate wikis, we argue that an effective strategy for lowering organizational costs is a mixed-initiative approach, by combining the unique abilities of humans for top-down organization of noisy data with the unique abilities of the machine to extract similar entities and manipulate content rapidly and at large scales.

In summary, we motivate our work by noting the high costs for organizing and maintaining content as one of the reasons why group wikis are currently underutilized, and present VisualWikiCurator as one way to lower these costs.

ACKNOWLEDGMENTS

We would like to thank our colleagues at XRCE and PARC (ASC).

REFERENCES

[1] Bier, E., Ishak, E., and Chi, E. Entity Workspace: an evidence file that aids memory, inference, and reading. Proc. of *ISI 2006*, 466--472, Springer-Verlag.

[2] Buffa, M. Intranet Wikis. In Proc. of the Intraweb Workshop, Proceedings of *WWW 2006*, ACM Press.

[3] Chan, B., Talbot, J., Wu, L., Sakunkoo, N., Cammarano, M., and Hanrahan, P. 2009. Vispedia: on-demand data integration for interactive visualization and exploration. In Proc.of *SIGMOD 2009*. ACM, NY.

[4] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proc. of *ACL'02*, Philadelphia, US.

[5] Grudin, J. and Poole, E. S. (2010). Wikis at work: success factors and challenges for sustainability of enterprise Wikis. Proc. of *Wikisyms 2009*, ACM, NY.

[6] Hanrahan, B., Convertino, G., Kong, N., Chi, E.H. (2011) Mail2Wiki: Posting and Curating Wiki Content from Email. In Proc. of ACM IUI 2011 conference.

[7] Happel H.J, Social search and need-driven knowledge sharing in Wikis with Woogle, In Proc. of *Wikisyms 2009*, ACM, New York, NY.

[8] Hoffart J., Zesch T. and Gurevych I. (2009). An architecture to support intelligent user interfaces for Wikis by means of Natural Language Processing, In Proc. of *Wikisyms 2009*, ACM, New York, NY.

[9] Hoffmann, R., Amershi, S., Patel, K., Wu, F., Fogarty, J., and Weld, D. S. 2009. Amplifying community content creation with mixed initiative information extraction. In *Proc. of* CHI '09. ACM, NY, 1849-1858.

[10] Holtzblatt, L. J., Damianos, L. E., and Weiss, D.. Factors impeding Wiki use in the enterprise: a case study. In Proc. of *CHI '10*, ACM, 4661-4676.

[11] Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In Proc. of *CHI '99*. ACM, NY, 159-166.

[12] Nakanishi T., Zettsu K., Kidawara Y., and Kiyoki Y., (2009) SAVVY Wiki: a context-oriented collaborative knowledge management system, In Proc. of *Wikisyms 2009*, ACM, New York, NY.

[13] Nardi, B. A., Miller, J. R., and Wright, D. J. (1998). Collaborative, programmable intelligent agents. *Commun. ACM* 41, 3, 96-104.

[14] Robertson S., Zaragoza H., Taylor M. Simple 2004 BM25 Extension to Multiple Weighted Fields. *CIKM'04*.

[15] Stasko, J., Görg, C., and Liu, Z. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization* 7, 2, 118-132