Selbstorganisierende Wikis

Prof. Dr. Iryna Gurevych, Technische Universität Darmstadt,
Torsten Zesch, Technische Universität Darmstadt[1]

Abstract. *Wikis sind kollaborativ genutzte, web-basierte Informationsspeicher, deren Bedeutung im Privat- und Arbeitsleben ständig steigt. Aufgrund ihrer leichten Benutzbarkeit wachsen Wikis oft schnell und ungesteuert, wodurch ihre Benutzbarkeit wieder dramatisch sinkt. Zur Lösung dieses Problems erforschen wir selbstorganisierende Wikis, die den Benutzer beim Einfügen und Suchen von Informationen unterstützen und ihn weitgehend von ablenkenden Routineaufgaben entlasten. Grundlage der Selbstorganisation sind erprobte Sprachverarbeitungstechniken wie Textgraphanalyse, Dokument-Clustering und Graph-basierte Termgewichtung. Das Forschungsprojekt untersucht dabei – neben anwendungsbezogenen technischen Aspekten – wie die Sprachverarbeitung eingesetzt werden kann, um die Schwelle der Mensch-Technik-Interaktion am Beispiel der Interaktion mit Wikis als neuartigem gemeinschaftlich genutzten Informationsspeicher, signifikant zu senken.*

1. Motivation

Wikis sind Software-Systeme, die es den Benutzern erlauben in einfacher Art und Weise Seiten im World Wide Web kollaborativ zu erstellen, zu verändern und zu verlinken [LC01]. Wikis finden zunehmend Eingang in verschiedenen Bereichen des Privat- und Arbeitslebens, z.B. über Wikipedia und Unternehmenswikis. Durch die niedrigen Hürden, die vor die Benutzung eines Wikis gestellt sind, wachsen Wikis meist schnell und ungesteuert. Damit wachsen aber auch wieder die Hürden, da das Finden von Informationen sowie das Hinzufügen von Informationen an einer geeigneten Stelle immer schwieriger werden.

Die zentrale Herausforderung im Zusammenhang mit kollaborativen Informationsspeichern wie Wikis ist es also, dem Benutzer einfaches Hinzufügen

[1] Ubiquitous Knowledge Processing Lab, Hochschulstr. 10, 64289 Darmstadt;
E-Mail: {gurevych,zesch} (at) tk.informatik.tu-darmstadt.de

ohne Beschränkungen zu erlauben und trotzdem das Wiki so zu organisieren, dass schnelles und zielgerichtetes Auffinden von Informationen möglich bleibt. Diese sich widersprechenden Zielsetzungen lassen sich durch Selbstorganisation des Wikis mittels Sprachverarbeitungstechniken vereinbaren. Ziel ist es, die Schwelle der Mensch-Technik-Interaktion signifikant zu senken.

2. Selbstorganisierende Wikis

Das Prinzip selbstorganisierender Wikis beruht auf der Anwendung von erprobten Techniken der automatischen Sprachverarbeitung auf die Inhalte eines Wikis. Die mit den momentan verfügbaren Sprachverarbeitungsalgorithmen erreichbare Ergebnisqualität reicht dabei aus, da wir Selbstorganisation nicht als automatischen, sondern als *semi-automatischen* Prozess verstehen. Der Benutzer soll ganz bewusst in die Entscheidungen mit einbezogen werden, da er nie das Gefühl haben darf, die Kontrolle über das System zu verlieren. Die Selbstorganisation des Wikis soll stets durch minimal ablenkende, vorschlagende Prozesse vom Benutzer gesteuert werden. Vollautomatisierung würde den Benutzer am Erforschen des Informations- und Wissensspeichers hindern. Hingegen soll das selbstorganisierende Wiki Benutzer aller Erfahrungsstufen von soviel unnötiger, ablenkender Arbeit wie möglich entlasten, wodurch ihre Aufmerksamkeit auf den Prozess der Wissenskonstruktion fokussiert bleiben kann.

Zum Erreichen dieser Ziele ist es notwendig, Techniken der Sprachverarbeitung zur:

- Selbststrukturierung des Wikis,
- verbesserten Suche im Wiki und
- Adaptierung der Benutzerschnittstelle des Wikis

einzusetzen.

2.1 Selbststrukturierung

Das Wiki-Paradigma erzwingt vom Benutzer keine strukturierte Vorgehensweise. Vielmehr wird angenommen, dass die Struktur während des Wachstums an die sich entwickelnden Bedürfnisse angepasst wird. Die notwendigen Restrukturierungen müssen dabei von den Benutzern selbst vorgenommen werden.

Werden z.B. in einem Unternehmen zwei Teams zusammengelegt, die jeweils eigene Bereiche im Wiki haben, müssen die Informationen geeignet

vereinigt werden. Da kein Team den Bereich des anderen Teams kennt, muss für jede Wiki-Seite aufwändig nach einem passenden Platz gesucht werden. Der Benutzer sollte hier nicht sich selbst überlassen bleiben, sondern das Wiki kann Vorschläge machen, an welcher Stelle eine Seite thematisch geeignet eingeordnet werden kann. Dazu müssen die Wiki-Seiten inhaltlich, sowie nach ihren Beziehungen untereinander [HMB07, ZG07], geclustert werden. Die Anzeige der Einfügevorschläge zieht Adaptionen der Benutzerschnittstelle nach sich, welche in Abschnitt 2.3 näher beschrieben werden.

Deutlich häufiger als das Zusammenlegen zweier Teams ist in der Praxis jedoch das Einfügen neuer Informationen. Es kann aber analog behandelt werden, da auch hier der Benutzer erst die passende Einfügeposition für eine Seite suchen müsste. Stattdessen kann er einfach die Wiki-Seite erstellen, anhand deren Inhalts das selbstorganisierende Wiki dann Vorschläge für Einfügepositionen errechnet, an denen die Seite thematisch kohärent eingeordnet werden kann.

Eine solche Selbststrukturierung verringert die Gefahr, dass Informationen mehrfach gespeichert werden, was zu hohen Kosten (5 Millionen Dollar jährlich in einem Unternehmen mit 1000 Mitarbeitern [FS03]) führen kann.

Informationen im Wiki werden jedoch erst besonders wertvoll durch die Beziehungen zu anderen Informationen, welche durch Links zu anderen Seiten ausgedrückt werden. Daher soll das Wiki automatisch passende Links zu anderen Wiki-Seiten vorschlagen. Die Vorschläge werden wir dabei mittels Sprachverarbeitungstechniken zur automatischen Verlinkung, basierend auf Schlagwortextraktion und Lesartendisambiguierung, generieren [MC07]. Die wichtigsten Schlagwörter können auch direkt als „Tags“ für die Seite vorgeschlagen werden. Ist eine Seite erst einmal in den Kontext anderer Seiten eingebettet, kann durch ‘Template Induction’ [IS07] das Layout der Seite an die umliegenden Seiten angepasst werden.

Semantische Wikis [KVV07] reichern die Beziehungen zwischen Wiki-Seiten durch explizite Semantik an. Zum Beispiel können Links zwischen den persönlichen Seiten von Wiki-Benutzern mit der Semantik „Mitarbeiter-von“ belegt werden. Daraus kann leicht eine automatische Hierarchie der Mitarbeiter im Unternehmen generiert werden.

Selbstorganisierende Wikis und semantische Wikis sind dabei orthogonal zueinander und können zu *semantischen selbstorganisierenden Wikis* kombiniert werden. Voraussetzung sind hinreichend genaue Algorithmen zur Relationsextraktion und Linkklassifikation [GD05], die im Moment nur begrenzt zur Verfügung stehen.

2.2 Verbesserte Suche

Eine verbesserte Suche bleibt trotz Selbststrukturierung notwendig, da eine Navigation zu relevanten Informationen in schnell wachsenden Wikis oft nicht mehr effizient möglich ist. Ein Unternehmen mit 1000 Mitarbeitern verliert jährlich bis zu 2,5 Millionen Dollar, weil ein Wissensarbeiter 15-25% der Arbeitszeit mit nicht-produktiver Informationssuche verschwendet [FS03].

Bei konventioneller Suche, die auf exaktem Vergleich der Stichwörter basiert, muss man das im Wiki verwendete Vokabular kennen. So bleibt eine Suche nach „Dienstreise“ erfolglos, wenn im Wiki konsistent „Geschäftsreise“ verwendet wird. Semantisches Information-Retrieval [GMZ07] löst dieses Problem, indem es Synonyme und andere verwandte Wörter in die Suche einbezieht.

Zur weiteren Verbesserung der Suche können mittels *Question-Answering* [SH08], Erkennung von Eigennamen [BP06], automatischer Indexgenerierung [CM07] und automatischer Erstellung domänenspezifischer Ontologien [MKT06] dem Benutzer fokussierte neue Suchpfade zur Verfügung gestellt werden.

Allerdings bleibt die Navigation in den Inhalten des Wikis eine dem Benutzer vertraute (vgl. ‚Surfen im Web’) und daher wichtige Technik. Deshalb sollte die Strukturierung des Wikis auch nicht durch einen Ansatz ersetzt werden, bei dem die Inhalte „flach“ organisiert und ausschließlich über die Suche zugänglich sind.

2.3 Adaptierte Benutzerschnittstelle

Im Verlauf des Selbststrukturierungsprozesses muss der Benutzer Entscheidungen treffen, die von der Benutzerschnittstelle eines selbstorganisierenden Wiki durch den Einsatz von Sprachverarbeitungstechniken unterstützt werden sollen. So muss z.B. beim Einfügen von Wiki-Seiten eine passende Position aus den vom Wiki generierten Vorschlägen ausgewählt werden. Die Benutzerschnittstelle sollte dazu jeweils einen Ausschnitt des Wikis an den vorgeschlagenen Einfügepunkten darstellen. Dabei kann neben dem manchmal wenig aussagekräftigen Seitentitel auch eine automatisch erstellte Textzusammenfassung [WB07] angezeigt werden.

Weitere Beispiele für die Adaptierung der Benutzerschnittstelle durch Techniken der Sprachverarbeitung sind:

- Wiki-Seiten, welche von den Benutzern als zu lang empfunden werden, können in Unterseiten zerlegt werden. Dies kann durch Textsegmentierung [F07] und die entsprechende Visualisierung der Grenzen in der Wiki-Seite unterstützt werden. Jede Unterseite wird dann wie beim Einfügen von neuen Inhalten der semi-automatischen Selbststrukturierung unterworfen und entsprechend eingeordnet.
- Die Markierung der Vertrauenswürdigkeit von Informationen in Wiki-Seiten kann durch *Trust-Coloring* [ABC07] auf Basis der Revisionshistorie des Wikis vorgenommen werden.
- Da Wiki-Seiten kollaborativ bearbeitet werden, ist in der visuellen Darstellung nicht mehr nachvollziehbar, welche Inhalte zu welchem Zeitpunkt von wem geändert wurden. Durch farbliche Hervorhebung der letzten Änderungen (sog. *Edit-Coloring*) kann die zum Erfassen der neuen relevanten Inhalte benötigte Zeit stark reduziert werden.
- Durch das dem selbstorganisierenden Wiki inhärente Clustering der Seiten, können auch leicht Empfehlungssysteme integriert werden, die den Benutzer unterstützen. So können z.B. alle Seiten angezeigt werden, die der aktuellen Seite ähnlich sind. Sind Benutzermodelle integriert, lässt sich auch Unterstützung der Art „andere Benutzer, die diese Seite anschauten, interessierten sich auch für folgende Seiten“ generieren.

3. Wiki-Mining

Für die Anwendung der in Abschnitt 2 vorgestellten Sprachverarbeitungstechniken auf ein Wiki reichen die explizit ausgezeichneten Informationen (z.B. Texte und einzelne Links) nicht aus, sondern es muss auch auf aggregiertes Wissen (wie z.B. über die Graphstrukturen des Wikis, die Gliederung der Seiten und die Benutzerinteraktionen) zugegriffen werden können. Dieses Wissen wird in einem *Wiki-Mining* Prozess [ZMG08] extrahiert. Dabei unterscheidet man zwischen:

- *Content-Mining* – dem Finden von relevanten Informationen in den Inhalten des Wikis
- *Structure-Mining* – dem Finden von Zusammenhängen zwischen Wiki-Seiten mittels der Linkstruktur, und
- *Usage-Mining* – dem Finden von relevanten Zusammenhängen in der Revisionshistorie und dem Nutzungsverhalten.

Spezielles Augenmerk liegt beim *Content-Mining* auf der Anpassung der Sprachverarbeitung an die Besonderheiten Benutzer-generierter Inhalte, z.B. Unvollständigkeit, Inkonsistenz, oder Fehleranfälligkeit.

Im selbstorganisierenden Wiki fließt das durch *Wiki-Mining* erzeugte erweiterte Wissen beim Einfügen und Suchen von Informationen semi-automatisch ein, wodurch die Strukturierung und Auffindbarkeit der Inhalte steigt. Dieser Informationskreislauf ist in Abbildung 1 dargestellt.

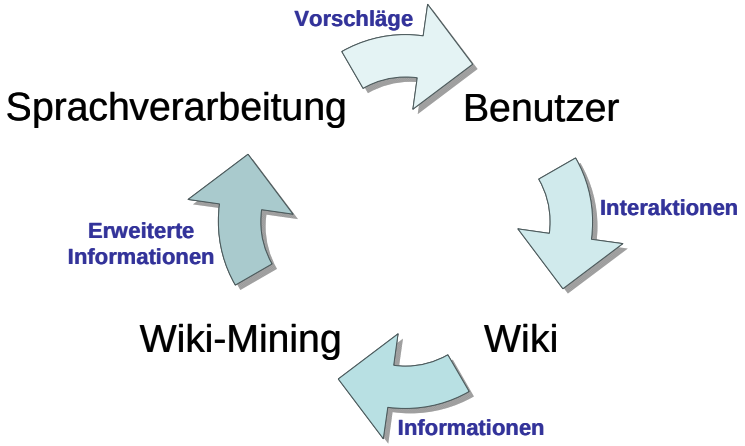

Abb. 1: Der Informationskreislauf im selbstorganisierenden Wiki.

3. Stand des Projektes

Das Forschungsprojekt „Selbstorganisierende Wikis“ wird am Ubiquitous Knowledge Processing (UKP) Lab der Technischen Universität Darmstadt durchgeführt. Das Projekt wird von der Klaus Tschira Stiftung gefördert und vereint die Expertise aus zwei Leuchtturm-Projekten des UKP Lab: „Wiki-Mining“ und „Darmstadt Knowledge Processing Repository (DKPro)“.

Das DKPro-Repository stellt eine Sammlung von skalierbaren, robusten und flexiblen Sprachverarbeitungskomponenten dar, basierend auf UIMA [FL04]. Das DKPro-Konzept zur Verarbeitung Benutzer-generierter Inhalte wurde 2007 mit dem von IBM weltweit vergebenen UIMA Innovation Award ausgezeichnet. Der Einsatz von UIMA in DKPro erlaubt die modulare Wiederverwendung von Komponenten und die schnelle Anpassbarkeit an neue Aufgaben, die im Kontext von selbstorganisierenden Wikis gelöst werden müssen.

4. Zusammenfassung

Das vorgestellte Forschungsprojekt untersucht selbstorganisierende Wikis mit dem Ziel, die Mensch-Technik-Interaktion beim Wissensmanagement signifikant zu erleichtern. Selbstorganisierende Wikis beruhen auf einem Informationskreislauf, bei dem durch Wiki-Mining zusätzliche Informationen aus dem Wiki gewonnen werden. Erprobte Algorithmen der Sprachverarbeitung benutzen diese Informationen zur Generierung von Strukturierungsvorschlägen, zur Verbesserung der Suche im Wiki und zur Adaptierung der Benutzerschnittstelle. Dadurch erleichtert das Wiki Benutzerinteraktionen, die den Informationskreislauf schließen und die Strukturierung und Auffindbarkeit der Wiki-Inhalte erhöhen.

Danksagung

Die Vorarbeiten wurden von der Deutschen Forschungsgemeinschaft im Projekt “Semantisches Information Retrieval” GU 798/1-2 & GU 798/1-3 gefördert. Das beschriebene Projekt „Selbstorganisierende Wikis“ wird von der Klaus Tschira Stiftung unter Projektnummer 00.133.2008 gefördert.

Literatur

[ABC07] Adler, B., Benterou, J., Chatterjee, K., de Alfaro, L., Pye, I., and Raman., V.: Assigning Trust To Wikipedia Content. Technical report, CSC-CRL-07-09, School of Engineering, University of California, Santa Cruz, CA, USA 2007

[BP06] Bunescu, R. and Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In Proceedings of EACL 2006. S. 9–16.

[CM07] Csomai, A. and Mihalcea, R.: Investigations in Unsupervised Back-of-the-Book Indexing. In Proceedings of FLAIRS 2007.

[F07] Ferret, O.: Finding Document Topics for Improving Topic Segmentation. In Proceedings of ACL 2007. S. 480–487.

[FL04] Ferrucci, D. and Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering, 10(3-4). 2004, S. 327–348.

[FS03] Feldman, S. and Sherman, C.: The High Cost of Not Finding Information. An IDC White Paper. 2007

[GD05] Getoor L., Diehl C.: Link Mining: A Survey. SigKDD Explorations Special Issue on Link Mining, 7:2, 2005.

[GMZ07] Gurevych, I., Müller, C., and Zesch, T.: What to be? - Electronic Career Guidance Based on Semantic Relatedness. In Proceedings of ACL 2007. S. 1032–1039.

[HMB07] Hassan, S., Mihalcea, R., and Banea, C.: Random-Walk Term Weighting for Improved Text Classification. In Proceedings of ICSC 2007. Irvine, CA.

[IS07] Irmak, U. and Suel, T.: Interactive Wrapper Generation with Minimal User Effort. In Proceedings of WWW 2007.

[KVV07] Krötzsch, M., Vrandecic, D., Völkel, M., Haller, H., Studer, R.: Semantic Wikipedia. Journal of Web Semantics, 5/2007, S. 251–261.

[LC01] Leuf, B. and Cunningham, W.: The Wiki Way: Quick Collaboration on the Web. Addison-Wesley, London. 2001.

[MC07] Mihalcea, R. and Csomai, A.: Wikify!: Linking Documents to Encyclopedic Knowledge. In Proceedings of CIKM 2007. S. 233–242.

[MKT06] Müller, S., Kritzler, N., Tartakovski, A., Bergmann, R., and Traphöner, R.: Knowledge Search within a Company-WIKI. In Proceedings of LWA 2006. S. 209-214

[SH08] Strzalkowski, T. and Harabagiu, S.: Advances in Open Domain Question Answering. Springer. 2008.

[WB07] Witte, R. and Bergler, S.: Next-Generation Summarization: Contrastive, Focused, and Update Summaries. In Proceedings of RANLP 2007. S. 27–29.

[ZG07] Zesch, T. and Gurevych, I.: Analysis of the Wikipedia Category Graph for NLP Applications. In Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007). S. 1–8.

[ZMG08] Zesch, T., Müller, C., and Gurevych, I.: Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In Proceedings of LREC 2008.