HiER|2013

Griesbaum, Heuwing, Ruppenhofer, Werner (Hrsg.)

HiER 2013

Proceedings des 8. Hildesheimer Evaluierungs- und Retrievalworkshop

Hildesheim, 25.–26. April 2013

J. Griesbaum, B. Heuwing, J. Ruppenhofer, K. Werner (Hrsg.):
HIER 2013 - Proceedings des 8. Hildesheimer Evaluierungs-
und Retrievalworkshop, Hildesheim 2013

© Institut für Informationswissenschaft und Sprachtechnologie,
Universität Hildesheim, 2013.

Inhaltsverzeichnis

Veranstalter & Tagungsteam 5

Workshopchairs 6

Programmkomitee 6

Vorwort 7

Session 1: Evaluierung im Information Retrieval 9

Martin Potthast, Matthias Hagen, Michael Völkse, Benno Stein
Crowdsourcing Interaction Logs to Understand Text Reuse from the Web 11

Melanie Dick
Einfluss morphologischer Werkzeuge auf die Information Retrieval-Qualität 17

Lukas Bittner
Konzeption eines Rechercheportals für die Musiktethnologie 23

Session 2: Sprachtechnologie für Sentimentanalyse und Opinion Mining 29

Jasper Brandes, Josef Ruppenhofer
Defining a gold standard for polar intensity odering 31

Michael Wiegand
Predicate Acquisition for Opinion Holder Extraction 41

Josef Ruppenhofer, Julia Maria Struß, Jonathan Sonntag, Stefan Gindl
IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches 51

Session 3: Mobile Informationsnutzung **59**

Jörg Cassens, Florian Scharf, Felix Schmitt, Michael Herczeg
What have they been up to in Lübeck recently? 61

Lars Wendt
Erfolgsfaktoren mobiler Anwendungen für klein- und mittelständische Unternehmen 67

Sören Dreyer
Gestaltungsrichtlinien für die Suche auf mobilen Webseiten 73

Session 4: Social Web **85**

Lisa Orszullok, Kathrin Knautz, Simone Soubusta
Aubruch nach Zyren 87

Clemens Roth, Joachim Griesbaum, Ralph Kölle
Was bedeutet 11.882 Bearbeitungen? 101

Katrin Weller, Axel Bruns
Das Spiel dauert 140 Zeichen 119

Veranstalter & Tagungsteam

Stiftung Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Universität Hildesheim
Bühler Campus, Gebäude Lübeckerstr. 3
31141 Hildesheim
www.uni-hildesheim.de/de/iwist.htm

Workshopchairs

Joachim Griesbaum, Ulrich Heid, Ben Heuwing, Thomas Mandl

Programmkomitee

Gertrud Faaß, Joachim Griesbaum, Ulrich Heid, Ben Heuwing, Ralph Kölle
Josef Ruppenhofer, Katrin Werner, Christa Womser-Hacker

Vorwort

Unser Umgang mit Information und den zur Verfügung stehenden Systemen ist im Umbruch: Wir wollen im Alltag überall auf kontextrelevante Informationen zurückgreifen, uns über aktuelle Entwicklungen in unseren sozialen Netzwerken informieren und mit den Systemen in natürlicher Sprache interagieren.

Der 8. Hildesheimer Evaluierungs- und Retrievalworkshop HIER 2013 thematisiert die Gestaltung und Evaluierung von Informationssystemen in diesen dynamischen Kontexten. Im Fokus stehen die folgenden Themen:

- Evaluierung im Information Retrieva
- Sprachtechnologie für Sentimentanalyse und Opinion Mining
- Mobile Informationsnutzung
- Social Web in der internen und externen Kommunikation von Organisationen

Dieser Band fasst die Vorträge des 8. Hildesheimer Evaluierungs- und Retrieval-Workshops (HIER) zusammen, der am 25. und 26. April 2013 an der Universität Hildesheim stattfand. Die HIER Workshop-Reihe begann im Jahr 2001 mit dem Ziel, die Forschungsergebnisse der Hildesheimer Informationswissenschaft zu präsentieren und zu diskutieren. Mittlerweile nehmen immer wieder Kooperationspartner von anderen Institutionen teil, was wir sehr begrüssen. HIER schafft auch ein Forum für Systemvorstellungen und praxisorientierte Beiträge.

Hildesheim, im April 2013
Joachim Griesbaum, Ben Heuwing, Josef Ruppenhofer, Katrin Werner

Session 1:

Evaluierung im Information Retrieval

Crowdsourcing Interaction Logs to Understand Text Reuse from the Web*

Martin Potthast, Matthias Hagen, Michael Völske, Benno Stein

Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract

We report on the construction of a new evaluation corpus for advanced research on text reuse. The corpus compiles manually written documents that are derived from a completely controlled, yet representative environment that emulates the web. Each of the corpus's 150 documents is about one of the 150 topics used at the TREC Web Tracks, thus forming a strong connection with existing evaluation efforts. Authors, hired at the crowdsourcing platform oDesk, had to retrieve sources for a given topic and to reuse text from what they found. Part of the corpus are detailed interaction logs that consistently cover the search for sources as well as the creation of documents. This will allow for in-depth analyses of how text is composed if an author is at liberty to reuse texts from a third party—a setting which has not been studied so far. In addition, the corpus will provide an original resource for the evaluation of text reuse and plagiarism detectors, a research field where currently only less realistic resources are employed. Altogether, the corpus's extent, its detail of logging, its relation to information retrieval, and, not least, its public availability sets it apart from existing resources.

* Extended abstract of an ACL 13 paper of the same title.

1 Introduction

The web has become one of the most common sources for text reuse. When reusing text from the web, humans often follow a three step approach: searching for appropriate sources on a given topic, copying of text from selected sources, and modifying or paraphrasing the reused text. A considerable body of research deals with the detection of text reuse, and, in particular, with the detection of cases of plagiarism.[1] Similarly, a large number of commercial software systems is being developed whose purpose is the detection of plagiarism. Both the manufacturers of these systems as well as researchers working on the subject matter frequently claim their approaches to be searching the entire web or, at least, to be scalable to web size. However, there is hardly evidence to substantiate this claim—rather the opposite can be observed: commercial plagiarism detectors have not been found to reliably identify plagiarism from the web (Köhler and Weber-Wulff, 2010), and the evaluation of research prototypes even under laboratory conditions shows that there is still a long way to go (Potthast et al., 2010). We explain the disappointing state of the art by the lack of realistic, large-scale evaluation resources.

With our work, we want to contribute to closing this gap. In this regard the paper in hand introduces a new corpus that, for the first time, emulates the entire process of reusing text from the web, both at scale and in a controlled environment. The corpus comprises a number of features that sets it apart from previous ones: (1) the topics of each document in the corpus are derived from the topics of the TREC Web Track, and sources have been retrieved manually from the ClueWeb corpus. (2) The search for sources is logged, including clickthrough and browsing data. (3) A fine-grained edit history has been recorded for each document. (4) A total of 150 documents were written each of which at least 5000 words long, whereas diversity is ensured via crowdsourcing. Altogether, this corpus forms the current most realistic sample of detailed interaction logs of authors reusing text. The corpus will be made publicly available.

1 The reuse of text with the intent of disguising the fact that text has been reused.

1.1 RelatedWork

Recently an evaluation framework has been introduced for the purpose of the annual PAN plagiarism detection competitions (Potthast et al., 2010). It provides, among others, a series of corpora that consist of automatically generated cases of plagiarism. Although this framework has been adopted by the community, a number of shortcomings render it less realistic:

1. All plagiarism cases were generated by randomly selecting text passages from documents and inserting them at random positions in another host document. This way, the reused passages do not match the topic of the host document.
2. The majority of the reused passages were modified in order to obfuscate the reuse. However, the modification strategies applied are, again, basically random: shuffling, replacing, inserting, or deleting words randomly. An effort was made to avoid non-readable text, but none of it bears intentional semantics.
3. The corpus documents are parts of books from the Project Gutenberg. Many of these books are rather old, whereas today the web is the predominant source for text reuse.

The authors of the PAN corpora tried to overcome the second issue: about 4000 passages were rewritten manually via crowdsourcing on Amazon's Mechanical Turk. However, a reused passage can still be spotted more easily by a topic drift analysis than by searching its original source, because of the first issue that the passage itself has been inserted at random (Potthast et al., 2011). From these observations it becomes clear that there are clear limits for the automatic construction of such kinds of corpora. The corpus presented in this paper addresses all of the mentioned issues since it has been constructed manually.

Besides the PAN corpora, there are only two other corpora that explicitly comprise reused text: the Clough09 corpus, and the Meter corpus. The Clough09 corpus consists of 57 short answers to one of five computer science questions which were reused from a respective Wikipedia article (Clough and Stevenson, 2011). While the text was genuinely written by a small number of volunteer students, the choice of topics is narrow, and text lengths range only from 200 to 300 words. Also, the sources from which to reuse text from were given up front, so that there is no data about the process

of how to retrieve such sources. The Meter corpus annotates 445 cases of text reuse in 1716 news articles (Clough et al., 2002). The cases of text reuse in this corpus are realistic for the news domain, however, they have not been created by the reuse process outlined. Note that in the news domain, text is often reused directly from a news wire without the need for retrieval. Our new corpus complements these two resources.

References

Barr, J., Cabrera, L. F. (2006). AI gets a Brain. Queue, 4(4):24–29, May.

Clough, P., Stevenson, M. (2011). Developing a Corpus of Plagiarised Short Answers. Lang. Resour. Eval., 45:5–24, March.

Clough, P., Gaizauskas, R., Piao, S. S. L., Wilks, Y. (2002). METER: MEasuring TExt Reuse. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 152–159, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cormack, G. V., Smucker, M. D., Clarke, C. L. A. (2011). Efficient and effective spam filtering and re-ranking for large web datasets. Information Retrieval, 14(5):441–465.

Elsayed, T., Lin, J. J., Metzler, D. (2011). When close enough is good enough: approximate positional indexes for efficient ranked retrieval. In Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011, pages 1993–1996.

Hiemstra, D., Hauff, C. (2010). MIREX: MapReduce information retrieval experiments. Technical Report TR-CTIT-10-15, University of Twente.

Jones, R., Klinkner, K. L. (2008). Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008, pages 699–708. ACM. 4 Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein

Köhler, K., Weber-Wulff, D. (2010). Plagiarism detection test 2010.

Pass, G., Chowdhury, A., Torgeson, C. (2006). A Picture of Search. In Xiaohua Jia, editor, Proceedings of the 1st International Conference on Scalable Information

Systems, Infoscale 2006, Hong Kong, May 30-June 1, 2006, volume 152 of ACM International Conference Proceeding Series, page 1. ACM.

Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. In Chu-Ren Huang and Dan Jurafsky, editors, 23rd International Conference on Computational Linguistics (COLING 10), pages 997–1005, Stroudsburg, Pennsylvania, August. Association for Computational Linguistics.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P. (2011). Overview of the 3rd International Competition on Plagiarism Detection. In Vivien Petras, Pamela Forner, and Paul D. Clough, editors, Working Notes Papers of the CLEF 2011 Evaluation Labs, September.

Potthast, M. (2011). Technologies for Reusing Text from the Web. Dissertation, Bauhaus-Universität Weimar, December.

Robertson, S.E., Zaragoza, H., Taylor, M.J. (2004). Simple BM25 extension to multiple weighted fields. In Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management, Washington, DC, USA, November 8-13, 2004, pages 42–49.

Viégas, F. B., Wattenberg, M., Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In CHI ’04: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 575–582, New York, NY, USA. ACM Press.

Einfluss morphologischer Werkzeuge auf die Information Retrieval-Qualität

am Beispiel des Porter-Stemmers und des Lemmatizers *wmtrans*

Melanie Dick

Universität Hildesheim
melaniedick@gmx.net

Zusammenfassung

Der Lemmatizer *wmtrans* der Canoo Engineering AG (Canoo 2012) und der Porter-Stemmer (Porter 1997) werden hier verglichen. Ziel ist es, jeweils den Einfluss auf die Information Retrieval-Qualität zu analysieren. Basierend auf den Testdaten der CLEF Studie von 2002 wird der Schwerpunkt auf die Analyse einzelner Informationsbedürfnisse gelegt.

Abstract

The lemmatizer *wmtrans* developed by Canoo Engineering AG (Canoo 2012) and the Porter stemmer (Porter 1997) will be compared regarding their influence on information retrieval quality. Based on test data from CLEF 2002.

Einleitung

In der Studie wurde der Lemmatizer der Firma Canoo Engineering AG (Canoo 2012) in Basel, welche diese Studie initiierte, mit dem Porter-Stemmer (Porter 1997) verglichen. Im Kontext des Information Retrieval sollte untersucht werden, ob und wie morphologische Werkzeuge die Treffergenauigkeit sowie die Vollständigkeit in Suchlösungen verbessern. Während der Lemmatizer auf lexikographischem Wissen basiert (also auf Wissen

über viele einzelne Wörter), führt ein Stemmer den Wortabgleich regelbasiert durch.

Methodik

Die CLEF-Studie von 2002 (CLEF 2003) wurde als Testdatensatz verwendet. Aus den 50 Informationsbedürfnissen dieser Studie wurden nach zuvor definierten Regeln ad-hoc-Anfragen generiert. Informationsbedürfnis 123 lautete zum Beispiel *„Heirat Jackson-Presley. Suche Dokumente, die über die vermutete Ehe von Michael Jackson mit Lisa Marie Presley oder ihre Trennung berichten“.* Hier wurden die sechs Begriffe *Heirat, Michael, Jackson, Lisa, Presley* und *Trennung* gewählt.

Komposita wie beispielsweise der Begriff *Literaturnobelpreisträger* von Informationsbedürfnis 94 wurden nach vier Varianten zerlegt. In zwei Varianten orientierte sich die Kompositazerlegung jeweils an dem Wissen der Wörterbücher canoo.net und duden.de. Die Version von duden.de wurde angelegt, um dem Aspekt, dass wmtrans durch die Wahl des eigenen Wörterbuchs einen Vorteil erhalten könnte zu entkräften. Der wmtrans-Lemmatizer basiert auf dem canoo.net-Wörterbuch. In einer dritten Version fand keine Wortzerlegung statt. In der Vierten wurden alle Kompositabildungen zerlegt. Der Suchbegriff *Literaturnobelpreisträger* wurde nach dem canoo.net-Wörterbuch in *Literaturnobelpreis* und *Träger* zerlegt. Duden.de kannte den Begriff. In dieser Version sowie in der Version ohne Kompositazerlegung wurde *Literaturnobelpreisträger* verwendet. In der letzten Variante wurde der Suchbegriff in *Literatur, Nobel, Preis* und *Träger* zerlegt.

Die Annäherung an eine authentische Nutzeranfrage war ein weiteres Kriterium. Dies wurde zum Beispiel bei der Anzahl der Suchterme berücksichtigt.

Es wurden fünf morphologische Werkzeuge getestet. Als Baseline wurde eine Suchlösung, welche über kein morphologisches Wissen verfügt, implementiert. Sie führt lediglich einen Abgleich der Schreibweise durch.

Ein weiteres Werkzeug war der regelbasierte Porter-Stemmer. Er basiert auf dem von Martin Porter entwickelten Stemming-Algorithmus (Porter 1997).

Der auf lexikalischem Wissen basierende Lemmatizer wmtrans, welcher auf das morphologische Wörterbuch canoo.net zurückgreift, wurde neben der Basisversion in zwei zusätzlichen Versionen durch Wortzerlegung (also Wortbildungsanalyse, zum Beispiel *Mitglied(er) – staat(en))* erweitert. Eine der zusätzlichen Version erlaubte die Wortzerlegung in der Suchanfrage und den Testdatensätzen, in der zweiten Version wurde die Wortzerlegung nur in den Testdaten und nicht in den Suchtermen eingesetzt.

Die fünf Suchwerkzeuge waren die sogenannte Baseline, der Porter-Stemmer, der wmtrans-Lemmatizer ohne die Fähigkeit der Wortzerlegung, ein wmtrans-Lemmatizer mit der Fähigkeit der Wortzerlegung in den Testdaten und ein wmtrans-Lemmatizer mit der Fähigkeit der Wortzerlegung in den Testdaten und Suchtermen.

Analyse

In der Gesamtübersicht aller Anfragen waren zunächst nur geringe Unterschiede zwischen den Suchwerkzeugen zu erkennen.

Die Recall-Werte der 50 Einzelanfragen aus CLEF variieren stark. Einzelanfragen mit Recall-Werten von 0 bis 1 sind enthalten. Interessant war neben der Gesamtübersicht deshalb vor allem die Analyse der einzelnen Informationsbedürfnisse. Bei Ergebnissen mit einem unterschiedlich hohen Recall zwischen Porter-Stemmer und *wmtrans*-Lemmatizer lassen sich die Unterschiede, die in der differenzierten morphologischen Verarbeitung begründet sind, schnell erkennen.

Anfragen mit einem identischen Recall wurden näher analysiert. Die Precision und insbesondere die Precision nach einer bestimmten Anzahl gefundener Dokumente, können ein Indikator dafür sein, dass eine genauere Untersuchung der morphologischen Verarbeitung an einer Stelle der Ergebnismenge lohnenswert ist. Vor- und Nachteile der beiden morphologischen Werkzeuge und ihre Auswirkung auf das Ranking eines Dokuments in der Ergebnismenge werden in der Evaluation behandelt. Anhand ausgewählter Querys werden verschiedene Suchterme und ihre unterschiedliche morphologische Verarbeitung erörtert. Nur von einer Suchlösung gefundene Suchbegriffe wirken sich auf das Ranking der Ergebnisse aus. Dies kann zu einer Verbesserung, aber auch zur Verschlechterung der Precision führen. In einem weiteren Schritt wurde der Einfluss der maschinellen Wortzerlegung, welche

in zwei Versionen des Lemmatizers integriert wurde, auf das Suchergebnis analysiert.

Beispiel: Informationsbedürfnis 101

Die exemplarische Erörterung an dem Informationsbedürfnis Nummer 101 aus der CLEF-Studie illustriert dies. Aus dem Informationsbedürfnis „*Zypern und die EU. Kann Zypern Mitglied der EU werden?2 w*urden die drei Suchbegriffe *Zypern*, *EU* und *Mitglied* extrahiert. Interessant ist hier vor allem die Verarbeitung der Begriffe *Mitglied* und *Zypern*. *Zypern* wird im Suchprozess als obligatorischer Begriff verwendet. Dies bedeutet, dass *Zypern* in einem Dokument enthalten sein muss, damit es als relevant von den morphologischen Werkzeugen kategorisiert werden darf und in die Ergebnismenge aufgenommen wird.

Die Baseline konnte lediglich einen Abgleich der identischen Schreibweise durchführen und deshalb nur genau die in der Suchanfrage verwendete Schreibweise, *Zypern* beziehungsweise *Mitglied*, erkennen. Der Recall ist bei der Baseline nicht vollständig.

Der Porter-Stemmer und der *wmtrans*-Lemmatizer konnte mit dem Suchbegriff *Mitglied* auch das Wort *Mitglieder* abgleichen. Die beiden um die Fähigkeit der Kompositazerlegung erweiterten Versionen waren auch in der Lage das Kompositum *Mitgliedsstaaten* zu erkennen. Die Version des *wmtrans*-Lemmatizers, welche um die Fähigkeit der Wortzerlegung in den Suchbegriffen erweitert wurde, zerlegt den Begriff *Mitglieder* in *Mit* und *glied*. *Mit* wurde allerdings als Stoppwort erkannt und deshalb im weiteren Suchprozess ignoriert.

Der Suchbegriff *Zypern* wurde vom Porter-Stemmer auf den Wortstamm *zyp-* reduziert und als Verb behandelt. Diese Wortverarbeitung erlaubt es dem Porter-Stemmer nicht mehr *Zyperns*, den Genitiv des Wortes, mit dem Suchbegriff *Zypern* abzugleichen, was dazu führt, dass fünf relevante Dokumente vom Porter-Stemmer und der Baseline nicht gefunden wurden.

Der Lemmatizer wmtrans konnte den Genitiv *Zyperns* und den Suchbegriff *Zypern* einander zuordnen. Auch das Adjektiv *zyprisch* konnte *Zypern* zugeordnet werden.

Die beiden um die Wortzerlegung erweiterten Versionen des Lemmatizers wmtrans konnten auch *Nordzypern* mit *Zypern* abgleichen. Dies war bei dieser Query allerdings unvorteilhaft, da es sich bei Texten mit diesem Kompo-

situm um ein anderes Thema handelte. Das Auffinden der Texte hatte einen negativen Einfluss auf die Precision, da hauptsächlich nicht relevante Texte gefunden wurden.

Fazit

Der Einfluss morphologischer Werkzeuge auf die Qualität des Information Retrieval wurde in dieser Studie an der Baseline, dem Porter-Stemmer und drei Varianten des *wmtrans*-Lemmatizers getestet. Zwei waren um die Fähigkeit der Kompositazerlegung erweitert. Als Baseline diente eine Version, die lediglich einen Wortabgleich identischer Wörter durchführt.

Die Qualität der Ergebnisse ist abhängig vom Informationsbedürfnis, beziehungsweise von der Komplexität der gewählten Suchbegriffe. Bei komplexen Suchbegriffen, wie zum Beispiel Komposita, sind die beiden *wmtrans*-Lemmatizer, welche um die Eigenschaft der Kompositazerlegung erweitert wurden, in den meisten Fällen überlegen. Die beiden um die Fähigkeit der Wortzerlegung erweiterten wmtrans-Lemmatizer verbesserten den Recall-Wert. Die Precision wurde allerdings verschlechtert.

Literatur

Canoo Engineering AG (2012). Produkt-Homepage: Canoo LanguageTool (ehemals wmtrans) 2012 – URL: http://www.canoo.com/languageexperts/products/toolkit/?l=de – Zugriffsdatum: 29.03.2013.

[CLEF 2003] BRASCHLER, Martin; GONZALO, Julio; KLUCK, Martin (2003)(Hrsg.). Advances in Cross-Language Information Retrieval. Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers. Springer Verlag, 2003.

DICK, Melanie (2012). Einfluss morphologischer Werkzeuge auf die Information-Retrieval-Qualität 2012 – Bachelorarbeit. Universität Hildesheim.

PORTER, Martin F. (1997). An algorithm for suffix stripping. In: SPARCK JONED, Karen (Hrsg.); WILLETT, Peter (Hrsg.): Readings in information retrieval. San

Francisco, CA, USA: Morgan Kaufmann Publisher Inc. 1997, S. 313-316. - URL: http://tartarus.org/martin/PorterStemmer/def.txt - Zugriffsdatum: 29.03.2013.

Konzeption eines Rechercheportals für die Musikethnologie

In Kooperation mit dem Center for World Music der Universität Hildesheim

Lukas Bittner

Universität Hildesheim, bittnerl@uni-hildesheim.de
Bahrfeldtstraße 9
31135 Hildesheim

Zusammenfassung

Dargestellt wird die Konzeption eines Portals für musikethnologische Recherche in Kooperation mit dem Center for World Music der Universität Hildesheim. Auf Basis eines vorangegangenen universitären Projekts, in dem ein analytischer Vergleich (Benchmark) bestehender Online-Kataloge durchgeführt und aus dessen Ergebnissen ein erster prototypischer Entwurf entwickelt wurde, knüpft die Arbeit (Bittner 2012) mit einer detaillierten Beschreibung des prototypischen Entwicklungsprozesses an. Darüber hinaus wird im Hinblick auf die durchgeführte Evaluierung das methodische Vorgehen beschrieben. Basierend auf der gewählten Erhebungsmethode des Experteninterviews stellt dieser Beitrag einen theoretischen Neuentwurf vor, in den die analytischen Ergebnisse einfließen.

Problemkontext

Das Center for World Music bildet seit dem Jahr 2009 eine weitere fachübergreifende Forschungseinrichtung der Universität Hildesheim. Mit dem Schwerpunkt der Musikethnologie versteht sich das Center als Schnittstelle von Wissenschaft und Öffentlichkeit. Den Grundstein für die Arbeit des CWM bilden umfangreiche Tonträger- und Instrumenten-Sammlungen. Die

größte Sammlung firmiert unter dem Namen „Music of Man Archive“ des Musikethnologen Prof. Dr. Wolfgang Laade. Diese umfasst etwa 46.000 Tonträger, 13.000 Bücher und 1.000 Instrumente, wird durch die Stiftung Niedersachen finanziert und als Dauerleihgabe der Universität in Hildesheim zur Verfügung gestellt. Da die Sammlungen in ihrer Varietät nicht nur gelagert, sondern der Öffentlichkeit strukturiert zugänglich gemacht werden sollen, befasst sich das CWM seit geraumer Zeit mit Erschließungsprojekten um die Artefakte sinnvoll zu präsentieren. Hierbei werden zunächst ausschließlich Materialien selektiert und zur weiteren Erschließung freigegeben, die für die Forschung in der Musikethnologie interessant sind und nicht in Bereiche wie beispielsweise der historischen Musikwissenschaft einzuordnen sind.

Standards für die Archivierung und die Katalogisierung von Musikarchiven in Europa wurden vorab durch ein EU-Projekt mit dem Namen DISMARC (Discovering Music Archives) festgelegt. Im Rahmen dieses Projekts entstanden unter anderem Vokabularien zur Verschlagwortung von Tonträgern. Diese Schlagwortliste wurde mangels Fachkräften und Expertise zunächst von dem CWM übernommen und zur Katalogisierung verwendet. In den kommenden Jahren soll jedoch an einer Verbesserung des Vokabulars gearbeitet werden, das die Bedürfnisse des Centers einbezieht.

Mit dem Sitz des CWM in der ehemaligen Timotheuskirche in Hildesheim, die unter anderem auch eine umfangreiche Instrumentensammlung des Privatsammlers Rolf Irle beheimatet, bietet sich neben dem Kirchenschiff auch im angrenzenden Pfarrhaus die Möglichkeit zur Lagerung der Artefakte. Zusätzlich zur Informationsbeschaffung über die wissenschaftliche Koordinatorin Frau Kerstin Klenke ergaben sich weitere interessante Kontaktmöglichkeiten zu Mitarbeitern des CWM. Durch das Treffen mit Frau Christine Weber, Diplom Bibliothekarin und Projektkoordinatorin für die Erschließung des Music of Man Archive, ergab sich die Möglichkeit den Standort der HMTMH (Hochschule für Musik, Theater und Medien Hannover), auf der Expo Plaza in Hannover zu besuchen, an dem die Katalogisierung der zuvor selektierten Tonträger stattfindet. An diesem Standort werden die Schallplatten und CDs in drei Bereichen bearbeitet:

- Alle Cover, Beihefte und Labels werden eingescannt (aktueller Stand: ca. 22.000LPs)
- Zu jedem Tonträger wird ein Datensatz im OPACangelegt
- Die Musik wird digitalisiert

Durch den zusätzlichen Standort der Expo Plaza ergibt sich eine weitere hilfreiche Möglichkeit die Materialien zu lagern (Raumfrage). Jeder Tonträger-Katalogdatensatz wird automatisch mit den schon vorhandenen Scans und Audio-Digitalisaten verlinkt und kann im OPAC über einen Medien-Player eingesehen werden. Alle digitalisierten und erschlossenen Daten sind im GBV nachgewiesen. Der OPAC des CWM bietet die Möglichkeit, auf Audiodateien und Scans zuzugreifen.

Ein Informationssystem zu entwickeln, das den Anforderungen aller Benutzer gerecht wird, stellt eine große Herausforderung an die Entwickler. Für eine solche Aufgabe ist es unabdingbar die Anforderungen aus Kunden- sowie Nutzersicht zu kennen, um eine Zielgruppe eingrenzen zu können.

Die zugrunde liegende Arbeit (Bittner 2012) versucht sich dieser Thematik anzunähern und die Anforderungen systematisch zu erfassen. Das Center for World Music widmet sich in seiner Arbeit der musikethnologischen Forschung. Ein Portal, das den Zugriff auf Artefakte aus diesem Bereich ermöglicht, spricht mit großer Wahrscheinlichkeit die Zielgruppe der Musikethnologen und Forscher dieses Fachgebietes an. Die Arbeit (Bittner 2012) ist im Wesentlichen dadurch charakterisiert, dass sie Komponenten aus Design und Informationswissenschaft vereint. Im Hinblick auf das vorangegangene Bachelor-Projekt wurden die Fragen geklärt, welche bestehenden Online- Kataloge Technologien bereitstellen, die sich im eigenen Entwurf als sinnvoll erweisen werden. Während der Entwicklung der prototypischen Entwürfe steht der Nutzer im Fokus. Innerhalb des gesamten Entwicklungsprozesses wurde darauf Wert gelegt, die angebotenen Funktionen in ihrer grafischen Beschaffenheit soweit zu reduzieren, dass der Nutzer nicht mit einer Flut an Informationen überfordert wird.

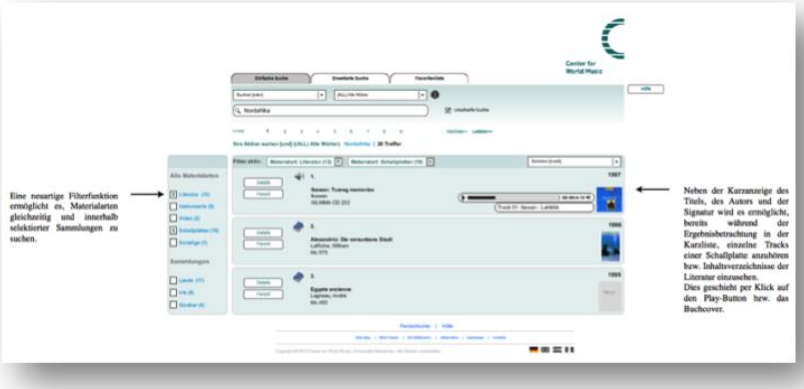

Abb.1: Kurzliste (Bittner 2012)

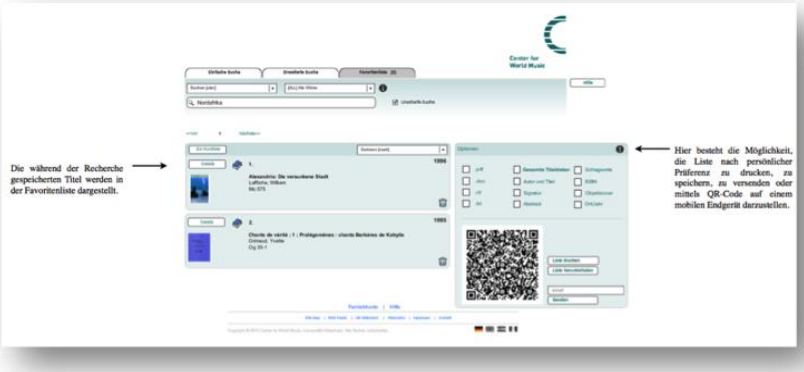

Abb.2: Favoritenliste (Bittner 2012)

Methodisches Vorgehen

Die Arbeit (Bittner 2012) beschreibt die Entwicklung eines prototypischen Entwurfs für musikethnologische Recherche für das Center for World Music der Universität Hildesheim. Dieser Entwurf wurde anhand von Experten, die aus den Bereichen der musikethnologischen Forschung, des Bibliothekswesens und des Museumswesens kommen, evaluiert. Durch dieses Vorhaben wurden konstruktive Meinungen zu dem Entwurf eingeholt, um diese in ei-

nem theoretischen Neuentwurf unterzubringen. Nach der einleitenden Einführung der Rahmenbedingungen und Zielsetzung (Bittner 2012) wird mit der Darstellung bestehender Kategorien der Prototypisierung angeknüpft. Diese werden angeführt, um die eigene prototypische Entwicklung thematisch einzugliedern. In einem weiteren Aspekt werden einerseits rückblickend auf das Bachelor-Projekt und die Entwicklung des MockUp mit dem Werkzeug „MockFlow“ eingegangen.

Anschließend an die einführenden Aspekte aus Bittner 2012, wird in einer detaillierten Beschreibung, die durch Grafiken aus dem Programm „MockFlow“ unterstützt werden, der Entwicklungsprozess der Entwürfe vorgestellt.

Bezüglich des methodischen Vorgehens werden sinnvolle Hinweise zur Methodik des Experteninterviews aus GLÄSER/LAUDEL (2006) herangezogen. Im Hinblick auf das leitfadenunterstützte Interview wird zudem auf die Entwicklung des Leitfadens eingegangen und erklärt, inwieweit dieser die weitere methodische Herangehensweise und Auswertung unterstützte, darüber hinaus wird eine Methodenreflexion zur Sprache kommen, in der die eigene Vorgehensweise kritisch analysiert wird.

Fazit und Ausblick

Der Schwerpunkt dieser Arbeit lag in der Beschaffung konstruktiver Meinungen zu den prototypischen Entwürfen. Besondere Anforderungen an die Entwürfe wurden bereits innerhalb der Arbeit am Bachelor-Projekt ermittelt. Neue Anforderungen ergeben sich nun zusätzlich aus den Experteninterviews. Diese weisen auf interessante Aspekte hin die in einem zukünftig geplanten Neuentwurf Erwähnung finden sollten. Die Ergebnisauswertung der Interviews zeigte, dass die entwickelten Funktionen, wie die Filterfunktion in der Kurzliste sowie die Zurverfügungstellung eines QR-Codes besonders wünschenswert im Hinblick auf eine geplante Umsetzung sind. Die ausgewählte Forschungsmethode des Experteninterviews hat sich folglich für die Evaluierung des während des Bachelor-Projekts entwickelten Prototypen als sinnvoll erwiesen. Die Umsetzung der gewonnenen Erkenntnisse in einem überarbeiteten Entwurf stellt einen wichtigen Fortschritt im Hinblick auf die reale Implementierung dar. Das Thema der Konzeption eines Portals für musikethnologische Recherche bietet auch für Studenten verschiedener Fachbe-

reiche und Studiengänge spannende Projektthemen. So besteht die Möglichkeit die Entwürfe nach Kriterien der Usability in Form von Benutzertests zu evaluieren. Darüber hinaus kann das komplette User Interface von Studenten der Designstudiengänge (bspw. Interaction Design, Webdesign etc.) überarbeitet und implementiert werden.

Eine Vision ist es Studenten aus diesen verschiedenen Fachbereichen zusammenzuführen und ein gemeinsames Projekt ins Leben zu rufen, um die Fertigstellung des Portals in geraumer Zeit gewährleisten zu können und das Center for World Music als internationale Forschungseinrichtung für Musikethnologie an die Öffentlichkeit zu kommunizieren.

Literaturverzeichnis

BERNHART, Mario; GRECHENIG, Thomas (2009): Softwaretechnik: Mit Fallbeispielen aus realen Entwicklungsprojekten. Pearson: München, S. 541.

BITTNER, Lukas (2012): Prototypisierung eines Portals für Musikrecherche und dessen Evaluation für das Center for World Music der Universität Hildesheim. Bachelor Thesis: Hildesheim

FLOYD, Christiane (1984): A Systematic Look at Prototyping. In: Budde, Reinhard; Kuhlenkamp, Karin; Mathiassen, Lars; Zullighoven, Heinz (Hrsg.): Approaches to Prototyping. Springer-Verlag: Heidelberg, S. 1-17.

HACKOS, JoAnn T.; REDISH, Janice C. (1998): User and task analysis for interface design. Wiley: New York, S. 378.

KALBACH, James (2008): Handbuch der Webnavigation. Die User-Erfahrung optimieren. O‘Reilly: Köln, S. 259.

KARLÖF, Bengt; ÖSTBLOM, Svantje (1994): Das benchmarking-Konzept. Wegweiser zur Spitzenleistung in Qualität und Produktivität. Vahlen: München.

KUNDLER, Wolfgang (2008): Magento. Erfolgreich mit dem ersten Online-Shop. Pearson: München, S.267.

PÜRER, Heinz (2003): Publizistik- und Kommunikationswissenschaft. Ein Handbuch. UVK: Konstanz, S. 541.

SCHNEIDT, Andreas (2007): Modellierung und Visualisierug kulturellen Erbes: Konzeption und prototypische Implementierung eines benutzerorientierten Web-Informationssystems für das Zentrum der Weltmusik in Hildesheim. Magisterarbeit: Hildesheim, S. 11.

Session 2:

Sprachtechnologie für Sentimentanalyse und Opinion Mining

Defining a gold standard for polar intensity ordering

Jasper Brandes, Josef Ruppenhofer

Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Marienburger Platz 22, 31141 Hildesheim
{jasper.brandes,josef.ruppenhofer}@uni-hildesheim.de

Abstract

In this paper, we report on an effort to develop a gold standard for the intensity ordering of subjective adjectives. Rather than pursue a complete order as produced by paying attention to the mean scores of human ratings only, we take into account to what extent assessors consistently rate pairs of adjectives relative to each other. We show that different available automatic methods for producing polar intensity scores produce results that correlate well with our gold standard, and discuss some conceptual questions surrounding the notion of polar intensity.

1 Introduction

As illustrated prominently by Hatzivassiloglou & Wiebe (2000), the correlation between gradability and evaluativeness can be used to harvest evaluative adjectives based on their occurrence in patterns that relate to gradability (e.g.*more/less entertaining*). The task that concerns us specifically is how to order gradable evaluative predicates that refer to the same scale relative to each other: how can one find out whether e.g. moronic is more negative than *stupid*?

In this paper, we address two aspects of the problem of intensity ordering. The first one is the fundamental question what a gold standard for this task should look like. Various researchers have proposed ways to automatically

assign valence scores to evaluative predicates. The methods used tend to produce distinct scores, and thereby absolute orderings, for all predicates. However, it is far from clear that humans would agree on complete orders between predicates such as *dumb, smart, stupid*. In fact, Kennedy & McNally (2005) suggest that by one set of linguistic criteria one can only distinguish between two major types of scale structures for adjectives: open and closed scales, depending on whether adjectives can be modified by intensifiers such *very* and *quite* (open scales) or by intensifiers such *utterly* and *absolutely* (closed scales).

We develop a gold standard that forms a sort of compromise between the complete orders that natural language processing techniques tend to produce and the bipartition that theoretical linguistics suggests. The second interest of our work is to compare how closely the results of different automatic methods of assigning intensity scores correlate with the human gold standard. Both our proposal for producing a gold standard and the evaluation of available methods are preliminary in that we have worked so far only with one scalar semantic domain, namely the degrees of intelligence associated with 18 English adjectives (see section 3 for a list).

2 Related Work

Acquiring knowledge about the polar strength of subjective expressions (*how positive/negative?*) has been pursued in three ways.

One obvious way is corpus-based, using the distributional properties of the scalar items to be scored or ranked in order to acquire intensity ratings. This approach has two important subtypes. In one, represented e.g. by Rill et al. (2012), scores for subjective expressions are derived from an extrinsic source, the star ratings metadata associated with reviews. In the second subtype, represented by Sheinman and Tokunaga (2009), language-intrinsic properties are used to assign intensity scores.

A second way of inducing scores or rankings consists in exploiting the knowledge inherent in lexical resources. SentiWordNet (Bacchianella et al. 2010), based on WordNet (Miller et al. 1990), is a prominent example of this approach. The main limitation of this approach is that it has to rely on the correctness and coverage of the existing resource.

A third way to acquire intensity ratings consists in re-purposing so-called affective norm data elicited from human subjects by psychologists. The elicitation usually involves the rating of individual words presented out of context. When done at large scale by way of crowd-sourcing to non-experts, as practiced by e.g. Warriner et al. (2013) using Amazon Mechanical Turk, affective norms promise good coverage. A practical limitation of this approach is, however, that no crowd may be available for the language of interest.

3 Constructing a Human Gold Standard

16 native speakers and 2 near-natives[1] were asked to assess the intensity of each of 18 intelligence-related adjectives in a fixed context. We used this setup so as to derive a ranking that would contain much less noise than automatic approaches, where factors like ambiguity and (lack of) coverage can play quite a substantial role. The adjectives we worked with are mainly culled from FrameNet's (Baker et al. 1998) MENTAL_PROPERTIES frame:

brainless, brainy, bright, brilliant, daft, dim, dimwitted, dumb, foolish, idiotic, imbecilic, inane, ingenious, intelligent, mindless, moronic, smart, stupid.

The survey was conducted online using a local installation of the Limesurvey[2] tool and consisted of three groups of questions. The first collected demographic information; the second elicited the ratings for the survey items (the adjectives); the third gave participants the chance to give feedback on the task. Ratings were elicited for each adjective individually, in randomized order, and under conditions intended to minimize bias. Specifically, to assess

1 The majority of the participants (n=9) claimed to speak American English (AE); 6 identified with the Australian variety (AuE) and 1 with New Zealand English (NZE).

2 www.limesurvey.org

the intensity of an adjective, the participant was asked to use a slider which could be moved from left to right by clicking the slider, dragging it in the desired direction and releasing the mouse click at the desired intensity. The slider scale ranged from -100 on the left to +100 on the right. The scores' meaning was described as follows: "Complete lack of intellectual characteristics equals -100, whereas extremely high intellectual characteristics equal +100." Participants were also given the following instruction on how to contextualize the meaning of the adjective: “Please rate the following adjective describing the intellectual characteristics of utterances and persons according to the intensity conveyed by it.”

The results of our data elicitation from 18 subjects are shown in Figure 1.[3] Two key aspects of the data are that a) each adjective has a different mean rating and b) the overlapping standard deviations suggest that not all adjectives can really be distinguished from each other with respect to intensity.

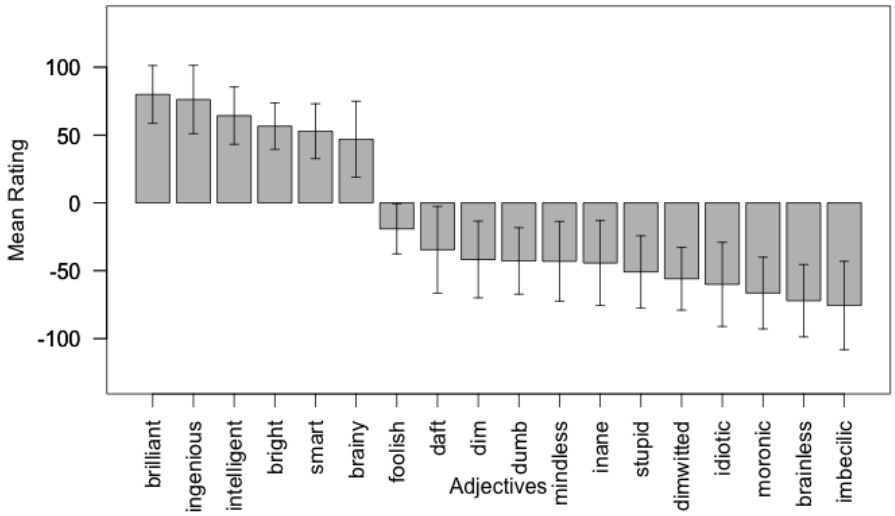

Figure 1 Human Evaluation: Means and Standard Deviations

3 24 subjects participated in the online survey. Three participants terminated the survey early, resulting in 21 full responses. Out of these 21 full responses, three were not included in the subsequent analysis. One response was a test run. For the two others it was evident from the ratings and the feedback on the survey that the participants had not entirely understood what they were expected to do.

In order to obtain some distinctions between, and grouping of, our adjectives by strength despite the overlapping standard deviations of the mean values, we decided to look at how consistently pairs of adjectives were rated across all participants.[4] Thus, for a given pair of adjectives we counted how many participants rated adjective A more intense than adjective B; adjective B more intense than adjective A; or adjective A equally intense as adjective B. With this data in hand, we defined a threshold of 12 out of the 18 raters who had to agree on an unequal rating so that we would adopt that as our relative ranking for the two adjectives.[5] Based on the predominant ordering relations for the pairwise comparison of the adjectives derived from the ratings for the individual adjectives, the six positive adjectives can be divided into three subgroups and the negative adjectives into 4, as shown in Figure 2.[6]

Intensity	**Subgroup of adjectives**
high positive	brilliant & ingenious
medium positive	intelligent
low positive	brainy, bright & smart
⇕	
low negative	daft & foolish
medium low negative	dim, dumb, inane, mindless & stupid
medium high negative	dim-witted & idiotic
high negative	brainless, imbecilic & moronic

Figure 2 Adjective Intensity Subgroups

4 Note that the comparison between two adjectives was not part of the survey but was constructed after the survey ended using simple combinatorics; only combinations of two adjectives with identical polarity needed to be investigated.

5 The number of participants needed for a qualified majority of raters is a tuneable parameter. In this first exploration, we decided to err on the side of being conservative and set it higher than the minimum number of 10.

6 In our data-set, we had no need to deal with any circular strength relations. But we feel that for wi*d*er application, we will need to be able to handle such cases.

4 A Comparison of Methods for Deriving Intensity scores

As discussed earlier, multiple methods for producing intensity scores have been developed. Here, we want to compare two of them relative to our gold standard for intelligence adjectives, namely one that relies on intensity scores derived extrinsically, from review metadata, and one that re-purposes ratings elicited by psychologists as affective norms.

4.1 Mean star ratings

The basic intuition of this method is that words in a review and the score of the review (or, its star rating) correlate: positive words are more likely to occur in positive reviews, and negative words more likely to occur in negative reviews. Thus, along the lines of Rill et al. (2012), we counted how many instances *n* of each adjective *i* of the set of intelligence adjectives occur in reviews with a given star rating *j* (score) in a corpus of Amazon reviews (Prettenhofer & Stein 2010).

$$SR_i = \frac{\sum_{j=1}^{n} S_j^i}{n}$$

Note that we slightly modify the approach of Rill et al. (2012): while they only considered the language in the review titles, we used only the body of the reviews. The reason for this is that we hope to benefit from a) greater coverage and b) more data points per adjective by using the body of the reviews rather than the titles.

4.1 Arousal scores

The affective ratings for almost 14,000 English words collected by Warriner et al. (2013) include scores of valence (from unhappy to happy), arousal (from calm to aroused) and dominance (from in control to controlled) for each word in the list. Scales from 1 to 9 were used for each of the ranges. This three-variable scoring system follows the dimensional theory of emotion by Osgood et al. (1957).

The scores for the words were elicited on the crowd-sourcing site Amazon Mechanical Turk from participants who needed to be US-residents. In each of the assignments (a set of about 350 words to be rated), the participants rated the words only along one of the three distinct dimensions (valence, arousal, dominance). Note that lemmas were assessed and not word senses. For our purposes, we will interpret the valence score as a relevant clue to the polar intensity score. The valence score was elicited as follows:

> You are invited to take part in the study that is investigating emotion, and concerns how people respond to different types of words. You will use a scale to rate how you felt while reading each word. There will be approximately 350 words. The scale ranges from 1 (happy) to 9 (unhappy).

As the focus on how *happy* the subjects *felt* indicates, the valence construct in theory is not defined identically to the notion of intensity we used since ours targets the degree of presence/lack of the underlying valued characteristic (intelligence).

4.3 Results

Table 1 shows the results of a Spearman rank correlation analysis between the gold standard (using the version in which some adjectives are grouped together into ordered ranks based on rater consistency), the mean star ratings and the affective norm valence scores, the latter two suitably transformed into ranks.[7]

Table 1: Correlations between intensities derived from different data

Data Sets	Correlation
Amazon Mean Star Ratings – Our data	0.837
Affective Norms Valence – Our data	0.673
Amazon Mean Star Rating – Affective Norms Valence	0.749

7 The correlation coefficient *rho* ranges between -1 and +1 for perfect negative and perfect positive correlation, respectively.

The results of the analysis show that both ways of deriving intensity scores produce good correlations with our human ratings. Interestingly, the reviews yield a better correlation than the affective norms. This is somewhat surprising in two respects. In the affective norms elicitation the relevance of the rating to specific lexical items was part of the design whereas it was indirect in the review data. Second, our use of occurrences in the text body involves an even more indirect relation between the star rating and the word occurrences than that used by Rill et al. (2012). Those authors focused specifically on occurrences of words in review titles rather than in the review text body based on the assumption that the words in the review titles would be more likely to have a clear connection to the star rating. Nonetheless, our overall results by using text bodies seem to be very good. In future work, we will seek to compare the usefulness of the body text of reviews relative to that of review titles by running parallel scoring experiments for multiple sets of adjectives from different semantic domains.

Finally, we also show the correlation between the valence scores from the affective norms dataset and the amazon star ratings. Considering this correlation is of interest for the following reason. Recall that the affective norms data elicitation used a different setup and asked specifically about how participants felt upon reading the word, whereas our instructions asked subjects to "[p]lease rate the following adjective describing the intellectual characteristics of utterances and persons according to the intensity conveyed by it". Our elicitation method may thus have focused more on the "objective" quality of intelligence rather than its evaluation. If that were so, then maybe Warriner et al.'s valence score should be considered the gold standard and our elicitation method one way of approximating it. On that understanding, our approach still fares well but worse than the approach that derives polar intensity scores from Amazon reviews (0.673 vs. 0.749).

Conclusion

In this paper, we presented an approach to deriving a human gold standard for polar intensity scores, using adjectives related to intelligence as our test case. Rather than producing a total order by only using the mean scores of our elicited responses, and rather than not making any distinctions due to the

relatively large standard deviations for all adjectives, we induce some subgroups by paying attention to the cases where a qualified majority of raters observe an asymmetric relation between the intensities associated with different adjectives.

We then used our gold standard to compare two ways of harvesting polar intensity scores. One uses review metadata and one interprets valence scores collected in a large affective norming experiment as polar intensity scores. Both methods and data sources produce good correlations with our gold standard, with the corpus-based approach being even better. What is attractive about the corpus-based approach is that it can run automatically and does not require the time and effort of elicitation. Nevertheless, since the review-based approach relies on a large review collection, such reviews need to be available in the first place. Further, and more significantly, there may be problems of coverage within collections of review data when one is interested in other domains or even general language vocabulary.

While the two approaches investigated here were shown to work for intelligence adjectives, we have not demonstrated their usefulness for other semantic domains. Further investigation of these and other methods for deriving polar intensity scores is needed.

Finally, we threw up a methodological and conceptual problem that is usually not directly addressed: when collecting human ratings for use as a gold standard, how should the task be stated to the participants? In most cases, having more of a desirable property correlates directly with more positive affect about the degree to which the property is exhibited. However, that need not be so: somebody described as a *know-it-all* in English or as *neunmalklug* in German may exhibit a very high degree of informedness but might still not be appreciated for it. For that reason, we will compare in future work for our intelligence adjectives and further sets of adjectives how well different ways of eliciting intensity information correlate.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA).

Baker, C.F., Fillmore, C.J. and Lowe, J.B. (1998) The Berkeley Framenet Project. In: Proceedings of the 17th international conference on Computational linguistics, Vol 1.86--90.

Hatzivassiloglou, V. and Wiebe, J. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In:Proceedings of the International Conference on Computational Linguistics (COLING-2000).

Kennedy, C. and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. Language, pages 345–381.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. International journal of lexicography, 3(4), 235-244.

Osgood, C.E., Suci, G.J. and Tannenbaum, P.H. (1957). The measurement of meaning. Vol. 47. Urbana: University of Illinois Press, 1957.

Prettenhofer, P. and Stein, B. (2010). Cross-Language Text Classification using Structural Correspondence Learning. In: 48th Annual Meeting of the Association of Computational Linguistics (ACL 10), 1118-1127, July 2010.

Rill, S., Scheidt, J., Drescher, J., Schütz, O., Reinel, D., and Wogenstein, F. (2012). A generic approach to generate opinion lists of phrases for opinion mining applications. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '12, pages 7:1–7:8. ACM.

Sheinman, V. and Tokunaga, T. (2009). AdjScales: Differentiating between Similar Adjectives for Language Learners. CSEDU (1): 229-235.

Warriner, A., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. http://crr.ugent.be/papers/Warriner_et_al_affective_ratings.pdf (accessed March 25, 2013).

Predicate Acquisition for Opinion Holder Extraction

A Data-Intensive Approach

Michael Wiegand

Spoken Language Systems, Saarland University
Gebäude C7 1
D-66123 Saarbrücken
michael.wiegand@lsv.uni-saarland.de

Abstract

Opinion holder extraction is one of the most important tasks in sentiment analysis. We will briefly outline the importance of predicates for this task and categorize them according to part of speech and according to which semantic role they select for the opinion holder. For many languages there do not exist semantic resources from which such predicates can be easily extracted. Therefore, we present alternative corpus-based methods to gain such predicates automatically, including the usage of prototypical opinion holders, i.e. common nouns, denoting for example experts or analysts, which describe particular groups of people whose profession or occupation is to form and express opinions towards specific items.

Introduction

Opinion holder extraction is one of the most important subtasks in sentiment analysis. It deals with the automatic extraction of entities uttering an opinion. For example, the opinion holders in Sentences (1) and (2) are *the vet* and *Russia*, respectively.

(1) The owner put down the animal although the vet had **forbidden** him to do so.
(2) Russia **favors** creation of "international instruments" to regulate emissions.

The extraction of sources of opinions is an essential component for complex real-life applications, such as opinion question answering systems or opinion summarization systems.

Most of recent research focused on supervised learning methods (Choi et al., 2005; Choi et al., 2006; Wiegand & Klakow, 2010). While such approaches typically yield robust classifiers, they rely heavily on a large amount of manually annotated sentences. Moreover, these methods are little telling with regard to the linguistic aspects underlying the classification task.

The Role of Predicates

By far the strongest clue for the task of opinion holder extraction are the predicates that syntactically relate to the candidate opinion holders (Wiegand & Klakow, 2011). The relevant types of predicates may be verbs (3), (deverbal) nouns (4) or adjectives (5).

(1) I **believe**$_{\mathbf{verb}}$ that this is more than that.
(2) This includes a growing **reluctance**$_{\mathbf{noun}}$ by foreign companies to invest in the region.
(3) Ordinary Venezuelans are even less **happy**$_{\mathbf{adj}}$ with the local oligarchic elite.

Moreover, the argument position in which the opinion holder is realized varies among the different lexical units. For verbs, the opinion holder may either be in agent position (6) or patient position (7). The former argument position, however, has been identified as – by far – the most frequently occurring construction (Bethard et al. (2004) report that 90% of the opinion holders on their datasets are agents.)

(4) China$_{agent}$ had always firmly **opposed** the US Taiwan Affairs Act.
(5) Washington **angered** Bejing$_{patient}$ last year.

Table 1 compares on the MPQA-corpus (Wiebe et al., 2005) a state-of-the-art supervised classifier that incorporates a large set of different features (the best classifier from Wiegand & Klakow (2010)) with a classifier that only contains knowledge of which predicates typically have an opinion holder as an argument. The predicate-based classifier should be considered an upper bound, since it already knows which predicates co-occur in principle with opinion holders and also in which argument position they can be found. (This information has been extracted from the labeled MPQA-corpus.) The table supports the usefulness of predicates for opinion holder extraction.

Table 1 Comparison of supervised and predicate-based classification.

Method	**Precision**	**Recall**	**F-Score**
Supervised Classifier	59.10	66.57	62.61
Predicate-based Classifier (Oracle)	47.04	68.62	55.82

Automatic Acquisition of Predicates

In a situation in which no labeled training data for supervised learning are available, a predicate-based classifier seems to be a promising solution. For the acquisition of the predicates, one can either employ manually compiled lexical resources, such as FrameNet (Baker et al., 1998) or Levin's verb classes (Levin, 1993) (this has been explored in Kim & Hovy (2006) and Wiegand & Klakow (2011)), or develop a data-driven induction method. In the following we will present an approach for the latter type of method. It is particularly attractive since it is only dependent on large amounts of unlabeled textual data rather than manually-compiled resources. As an unlabeled text corpus for this kind of induction, we make use of the North American News Text Corpus (LDC95T21).

Prototypical Opinion Holders

Even though an unlabeled corpus does not contain any explicit information about which entities are actual opinion holders, there exists a set of common nouns denoting particular groups of people who, since their profession or occupation is to form and express some opinion towards specific items, disproportionately often represent actual opinion holders. Therefore, we use the mentions of these nouns as a proxy for actual opinion holders. Typical examples of such prototypical opinion holders (protoOHs) are *advocates*, *analysts*, *critics*, *experts*, *reviewers* or *supporters*. Sentences (8) and (9) illustrate two mentions of such nouns.

(6) Experts$_{\text{protoOH}}$ **agree** it generally is a good idea to follow the manufacturer's age recommendation.
(7) Shares of Lotus Development Corp. dropped sharply after analysts$_{\text{protoOH}}$ **expressed** concern about their business.

Table 2 illustrates the verbs most highly correlating with protoOHs (we use *Pointwise Mutual Information*). All of these predicates have the opinion holder in agent position[1]. Intuitively, the verbs that are thus gained are plausible predictors for opinion holders.

Table 2 List of verbs most highly correlating with protoOHs.

accuse acknowledge anticipate argue agree attribute believe call caution charge change cite claim concede criticize conclude contend consider complain describe disagree doubt fear expect estimate find forecast hope look note praise predict point question recommend say see speculate suggest tell testify think try view warn write wonder worry

One of the advantages of the usage of prototypical opinion holders is that predicates with parts of speech other than verbs can also be extracted.

1 Semantic roles were recognized by a mapping from grammatical functions (e.g. nsubj(complain-v, proponents-n) typically represents a noun in agent position while dobj(anger-v, fans-n) a noun in patient position).

Predicates that Select Opinion Holders in Patient Position

The downside of protoOHs as a means to extract predictive predicates for opinion holder extraction is that it only reliably detects predicates having the opinion holder in agent position. If a protoOH occurs in patient position, it often does not function as an opinion holder (10).

(8) They **criticized**$_{\mathbf{verb}}$ their opponents$_{\text{protoOH}}$.

For that reason, we need to devise a complementary induction method to extract predicates selecting opinion holders in patient position.

Our induction method for these verbal predicates rests on the observation that the past participle of those verbs, such as *shocked* in Sentence (11), is very often identical to some predicative adjective (12) having a similar if not identical meaning. (For the predicative adjective, the opinion holder is, however, its agent and not its patient.)

(9) He had **shocked**$_{\mathbf{verb}}$ me. (opinion holder: patient)
(10) I was **shocked**$_{\mathbf{adj}}$. (opinion holder: agent)

Instead of extracting those verbs directly (11), we take the detour via their corresponding predicative adjectives (12). This means that we collect all those verbs from our unlabeled corpus for which there is a predicative adjective that coincides with the past participle of the verb. Table 3 shows some verbs that could thus be extracted.

Table 3 Examples of induced predicates that take opinion holders in patient position.

anguish astonish astound concern convince daze delight disenchant disappoint displease disgust disillusion dissatisfy distress embitter enamor engross enrage entangle excite fatigue flatter fluster flummox frazzle hook humiliate incapacitate incense interest irritate obsess outrage perturb petrify sadden sedate shock stun tether trouble

Evaluation

For the evaluation, we consider two datasets belonging to two different domains. Cross-domain evaluation is important in order to obtain a realistic view of the effectiveness of different classification methods. The first domain that we consider, ETHICS (5700 sentences), is the largest domain subset of the MPQA-corpus (Wiebe et al., 2005). It contains news items that discuss human rights issues (mostly in the context of combating global terrorism). The second domain, FICTION (614 sentences), is a set of summaries of fictional work (two Shakespeare plays and one novel by Jane Austen). We will compare both predicate-based and supervised classifiers. As the second domain corpus is much smaller than the first, it can only be used as a test corpus. We use ETHICS for in-domain evaluation for supervised classifiers (we carry out 5-fold cross validation) while FICTION is used as an out-of-domain dataset (i.e. we train on ETHICS and test on FICTION).

Another difference between the two domain corpora lies in the distribution of opinion holders in patient position as stated in Table 4. The proportion is notably higher on the out-of-domain dataset than on the in-domain dataset. This raises the question whether a supervised classifier trained on ETHICS with its very few opinion holders in patient position is able to identify all types of opinion holders on FICTION.

Table 4 Proportion of opinion holder as patients.

ETHICS	**FICTION**
1.47	11.59

Table 5 compares predicate-based classification with supervised classification. For supervised classification, we use the best classifier from Wiegand & Klakow (2010) that incorporates no external lexicons. For predicate-based classification we consider an automatic induction method (as presented in the previous section) and a classifier using manually compiled resources. The resources comprise *strong subjective expressions* of the Subjectivity Lexicon (Wilson et al., 2005) with a subset of Levin's verb classes (Levin, 1993). The latter lexicon is specifically designed to identify opinion holders in patient position. Our choice of manually-compiled lexicons was shown to be most

effective on a large-scale analysis of different resources (Wiegand & Klakow, 2011). Note that for each predicate-based classification method (i.e. *Induction* and *Lexicon*), we consider two configurations: AG just detects opinion holders in agent position while AG+PT detects opinion holders in both agent and patient position.

Table 5 shows that the supervised classifier only outperforms the predicate-based classifier on the in-domain evaluation. On the out-of-domain dataset *Lexicon* is notably better while *Induction* is only slightly worse than the supervised classifier. The supervised classifier is much worse on the out-of-domain dataset since supervised classifiers are very susceptible to domain mismatches. That classifier fails to correctly identify many opinion holders in patient position since only few of them have been included in the training set (i.e. ETHICS corpus). *Induction* is the weakest performing classifier but considering the little manual effort that is required to build it, its performance on ETHICS in comparison to *Lexicon* is already fairly remarkable.

The configurations of the predicate-based classifiers that also detect opinion holders in patient position are only effective on FICTION which can be explained by the fact that this dataset contains significantly more of such opinion holders than ETHICS (see Table 4). Since we observe an improvement on FICTION with *Induction* from AG to AG+PT, we can also conclude that our induction method to extract predicates that take opinion holders in patient positions works.

The gap between *Induction* and *Lexicon* is much larger on FICTION. This can be explained by the fact that *Induction* has a bias towards news texts (hence the good performance on ETHICS) since it was generated with the help of a large news corpus.

Table 5 Comparison of different classifiers (evaluation measure: F-Score).

	Predicate-based classifier				**Supervised classifier**
	Induction		**(Manually-compiled) Lexicon**		
Domain	**AG**	**AG+PT**	**AG**	**AG+PT**	**(No lexicons included)**
ETHICS (in-domain)	50.77	50.99	52.22	52.27	59.52
FICTION (out-of domain)	46.59	49.97	54.84	59.35	51.21

Finally, Table 6 examines whether the supervised classifier can be improved if information from either of the predicate-based classifiers is added. The table clearly shows that on both domains the addition of either of the predicate-based classifiers results in some improvement. However, the gain in improvement is greater on the out-of-domain dataset from which we conclude that the predicate-based classifier is an important means to bridge a domain mismatch between source and target corpus. Although *Induction* itself has a domain bias, this method is less restrictive than supervised classification. This can partly be explained by the fact that the amount of unlabeled data that *Induction* uses is much larger than the amount of labeled training data that is available for supervised learning.

Table 6 Combining the supervised classifier with predicate-based classifiers (evaluation measure: F-Score).

	Supervised classifier	**Supervised classifier +Induction**	**Supervised classifier+Lexicon**
ETHICS (in-domain)	59.52	60.42	61.50
FICTION (out-of domain)	51.21	59.58	65.67

Conclusion

We presented a data-driven approach for opinion holder extraction that focuses on the extraction of predictive predicates. Two induction approaches were examined, one relying on prototypical opinion holders and another that exploits the similarity between predicative adjectives and verbs. The former is designed to extract opinion holders in agent position (which is also the most frequently occurring argument position in general), while the latter is specifically designed for the extraction of predicates that take opinion holders in patient position. The induction methods were compared with a predicate-based approach using a manually compiled lexicon and supervised learning. The results are promising and suggest particular significance of predicate-based classification for cross-domain opinion holder extraction.

References

Baker, C., Fillmore, C., and Lowe, J. (1998). The Berkeley FrameNet Project. In *COLING*.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. 2004. Extracting Opinion Propositions and Opinion Holders using Lexical Clues. In Shanahan, J., Qu, Y., and Wiebe, J. (Eds.) *Computing Attitude and Affect in Text: Theory and Applications*.

Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *HLT/EMNLP*.

Choi, Y., Breck, E., and Cardie, C. (2006). Joint Extraction of Entities and Relations for Opinion Recognition. In *EMNLP*.

Kim, S., and Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In *ACL-Workshop on Sentiment and Subjectivity in Text*.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating Expressions and Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2/3): 164-210.

Wiegand, M., and Klakow, D. (2010). Convolution Kernels for Opinion Holder Extraction. In *HLT/NAACL*.

Wiegand, M., and Klakow, D. (2011). The Role of Predicates in Opinion Holder Extraction. In *RANLP-Workshop on Information Extraction and Knowledge Acquisition*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In *HLT/EMNLP*.

IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches

Josef Ruppenhofer[1], *Julia Maria Struß*[1], *Jonathan Sonntag*[2], *Stefan Gindl*[3]

[1]University of Hildesheim
Department of Information Science and Natural Language Processing
Marienburger Platz 22, 31141 Hildesheim
{josef.ruppenhofer | julia.struss} @ uni-hildesheim.de

[2]University of Potsdam
Computational Linguistics
Karl-Liebknecht-Straße 23-24, 14476 Potsdam
jsonntag @ uni-potsdam.de

[3]MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1, 1190 Wien
stefan.gindl @ modul.ac.at

Abstract

In this paper, we report on the definition of a shared task considering source (whose opinion?) and target (about what?) extraction in protocols of the Swiss parliament that will be conducted by the Interest Group on German Sentiment Analysis (IGGSA)[1].

1 http://iggsa.sentimental.li/

1 Introduction

In fine-grained opinion mining/sentiment analysis, the source or holder of an opinion as well as the target or object of opinion often need to be extracted from the text rather than being available from metadata. This is the case especially for text types that contain multiple sources and targets. While it may appear that sources and targets should be intuitively and reliably recognizable to humans, and we should thus find high inter-rater agreement among annotators, actual attempts at defining and annotating the concepts show that this is not so (see e. g. NTCIR-6: OAT (Seki et al. 2007)).

We report on the definition of a shared task related to source and target extraction in protocols of the Swiss parliament that the IGGSA is in the process of preparing and that is to be held in the run-up to the KONVENS conference in 2014. We discuss various ways of defining the notions source and target and report on the development of the annotation scheme.

2 Related Work

Sentiment analysis systems must be able to reliably tie opinions or subjective states to their sources and targets. This is a non-trivial task as some sentiment-bearing expressions are not linked to the sources, and some not even to the targets, of opinion. In the best case, the source and target correspond to semantic roles of sentiment-bearing predicates that can be expressed as syntactic arguments (cf. Ruppenhofer et al. 2008). However, a direct tie-in with semantic role labeling is usually not the chosen way of handling the source and target extraction. In what follows, we discuss the reasons for this and some of the adopted alternative problem statements.

In the case of one important sub-class of sentiment-bearing expressions, called expressive subjective elements by Wiebe et al. (2005), a grammatical link exists only between the opinion expression and the target, but not to the source. For instance, in the case of *idiotic* we know that what the adjective modifies or is predicated of is the target of the sentiment conveyed. Note, however, that the sources may differ between the two following examples: in

(1), the source is the writer of the text, whereas in (2) it is the quoted speaker Irvine.

(1) His rude, crude response and IDIOTIC exit from his duties is hardly deserving of the praise he has attracted. [www]

(2) ”That was IDIOTIC,” Irvine told talkSPORT. [www]

Rather than connect expressions of opinion only to their immediate sources, it is desirable to keep track of the chain of transmission. In the MPQA-corpus (Wiebe et al. 2005), levels of nesting are recorded that would show for a sentence like (2) that not only is Irvine the source of the opinion but also the writer of the text.

The main issue regarding targets is whether the analysis should address only what one may call “local” targets – expressions that are semantic valents and syntactic dependents of a particular sentiment-bearing predicate – or whether other pragmatically relevant targets should also be taken into account. To illustrate the difference, consider the following pair of examples:

(3) a. I am not a Dortmund fan - I am a Schalke fan - but I am GLAD + [Dortmund beat Bayern TARGET].

b. I am not a Dortmund fan - I am a Schalke fan - but I am GLAD Dortmund beat - [Bayern TARGET].

(3a) displays the stable, “literal” sentiment that is conveyed by the sentence: that the speaker is glad about the reported event. (3b), by contrast, displays an inferred sentiment: that the speaker specifically dislikes Bayern's team.

Very much a pragmatic understanding of targets is adopted by Stoyanov and Cardie (2008). They suggest a definition of opinion topic and present an algorithm for opinion topic identification that casts the task as a problem in topic co-reference resolution. They distinguish between *Topic* (subject of the opinion as intended by the opinion holder), *Topic span* (minimal span of text mentioning the topic) and *Target span* (syntactic surface form comprising the contents of the opinion). Example (4) gives an example of Stoyanov and Cardie's, showing through the question marks that there are multiple entities that might be perceived to refer to a relevant topic. Notice that *think* is treated as the relevant subjective expression.

(4) [OPINION HOLDER Al] thinks that [TARGET SPAN [TOPIC SPAN? the government] should [TOPIC SPAN? tax gas] more in order to [TOPIC SPAN? curb [TOPIC SPAN? CO2 emissions]]].

Thus, what Stoyanov and Cardie are after are certain inferred targets which may be more important on the text-level than the overt target and which may not have syntactic relations to the subjective expression.

In terms of prior shared tasks in opinion mining, the most relevant work to ours was done in the context of the Japanese NTCIR[2] Project. In the NTCIR-6 Opinion Analysis Pilot Task (cf. Seki et al. 2007), offered for Chinese, Japanese and English, sources and targets had to be found relative to whole opinionated sentences. If multiple opinions were expressed in a given sentence, the task allowed for multiple opinion holders, possibly occurring anywhere in the document. In the evaluation, where necessary, co-reference information was used to (manually) check whether a system response was part of the correct referent's chain of mentions. The sentences in the document were judged as either relevant or non-relevant to the topic (=target). Polarity was determined for each opinionated sentence and in case of multiple opinions, the polarity of the main opinion expressed was chosen.

All sentences were annotated by three assessors, allowing for strict and lenient (by majority vote) evaluation. The successor tasks, NTCIR-7 and -8: Multilingual Opinion Analysis, (Seki et al. 2008, 2010) were basically similar in their setup.

While the STEPS-task will focus on German, the most important difference to the shared tasks organized by NTCIR is that it defines the source and target extraction task at the level of individual subjective expressions. A key aspect that is yet to be resolved in the development of the STEPS-TASK is what kinds of units may be considered as the relevant subjective expressions, as discussed further below in section 4.

2 NII [National Institute of Informatics] Test Collection for IR Systems

3 Data

The data set comes from the Swiss parliament[3]. The choice of this particular data set is motivated as follows. On a pragmatic level, it was important that the source data along with our annotations could be redistributed freely. In terms of the task itself, we were interested in having a text type in which multiple sources and targets would be relevant. Further, several IGGSA-members are interested in, and collaborating with researchers in, political science in the context of the digital humanities.

The data set does have some special characteristics, though, when compared to other German-language data sets. First, it reflects the Swiss variety of German. Second, the Swiss parliament is a body that operates multilingually. In order to keep the influence of these characteristics to a minimum, only speeches in German discussing foreign affairs are selected for the tasks' gold standard.

4 Development of the annotation scheme

In order to meet different research interests of the IGGSA-members, a new annotation scheme is being developed based on the annotations of a first explorative annotation step. In this step 50 sentences are annotated by four independent annotators with respect to opinions, targets and sources with the only requirement, that sources and targets are to be annotated at the level of individual subjective expressions and all nested targets and holders have to be considered. The annotators are asked to comment their annotation decisions, which will then be used for the development of an initial annotation scheme that will be validated in a second annotation step with different annotators on the same data, before the actual annotation of the gold standard sentences will be conducted.

3 Schweizer Bundesversammlung: http://www.parlament.ch/ab/frameset/d/index.htm

While we want the task to be focused on the extraction of sources and targets, it is also crucial to have good agreement on subjective expressions since source and targets exist only relative to them. Consider example (5), slightly modified from the MPQA.

(5) The United States has [CONSISTENTLY ATTACHED] [GREAT IMPORTANCE] to this issue.

The MPQA annotations treat *consistently attached* as one subjective expression (a direct subjective) and *great importance* as another (an expressive subjective element). If one accepts this analysis, what should the sources and targets be for these two expressions? An alternative analysis that fuses the two subjective expressions into one, recognizing that *attach* is a support verb for *importance*, would avoid the problem of what to treat as the target of *attach*. Other problems may result when the (relative) lack of constraints on what may count as a subjective expression leads to a situation where the subjective expression overlaps the target.

5 Task description

We plan on offering one main task and two subtasks in the context of STEPS. In the main task the participants are asked to identify subjective expressions *and* their respective opinion holders and opinion targets. In the subtasks, these two aspects are divorced. For subtask 1 the participants are given the subjective expressions and are only asked to identify opinion holders and/or opinion targets. Subtask 2 is considered with the identification of subjective expressions and their polarities and strength. Participating in the main task precludes participation in subtask 1 and vice versa.

6 Conclusion

In this paper we discussed various ways of defining opinion targets and holders with more semantic/syntactic ways on the one hand and more pragmatic

ways on the other and reported on the task definition as wells as first ideas for the development of a new annotation scheme to be employed in the development of the gold standard for the shared task that will be held in the run-up of the KONVENS conference in 2014.

References

Ruppenhofer, J.; Somasundaran, S.; Wiebe, J. (2008): Finding the Sources and Targets of Subjective Expressions. In: Proceedings of the LREC, Marrakech, Morocco, 26 May - 1 June 2008, pp. 2781–2788.

Seki, Y.; Evans, D.; Ku, L.; Chen, H.; Kando, N.; Lin, C. (2007): Overview of Opinion Analysis Pilot Task at NTCIR-6. In: Proceedings of the 6th NTCIR. Tokyo, Japan, May 15-18, 2007, pp. 265–278.

Seki, Y.; Evans, D.; Ku, L..; Le Sun; Chen, H.; Kando, Noriko (2008): Overview of Multilingual Opinion Analysis Task at NTCIR-7. In: Proceedings of the 7th NTCIR, Tokyo, Japan, December 16-19, 2008, pp. 185–203.

Seki, Y.; Ku, L.; Sun, L.; Chen, H.; Kando, N. (2010): Overview of Multilingual Opinion Analysis Task at NTCIR-8. A Step Toward Cross Lingual Opinion Analysis. In: Proceedings of the 8th NTCIR. Kyoto, Japan, June 15-18, 2010, pp. 209–220.

Stoyanov, V.; Cardie, C. (2008): Topic identification for fine-grained opinion analysis. In: Proceedings of the 22nd COLING. Manchester, UK, 18-22 August 2008, pp. 817–824.

Wiebe, J.; Wilson, T.; Cardie, C. (2005): Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation* 39 (2), pp. 165–210.

Session 3:

Mobile Informationsnutzung

What have they been up to in Lübeck recently?

Jörg Cassens[1], *Florian Scharf*[2], *Felix Schmitt*[2], *Michael Herczeg*[2]

[1]University of Hildesheim
Institute for Mathematics and Applied Informatics
cassens@cs.uni-hildesheim.de

[2]University of Lübeck
Institute for Multimedia and Interactive Systems
[scharf|schmitt|herczeg]@imis.uni-luebeck.de

Abstract

This talk will give an overview over three related research prototypes for ambient interactive systems. We start by introducing NEMO, the Network Environment for Multimedia Objects. NEMO is a smart media environment for semantically rich, personalised, and device-specific access to and interaction with multimedia objects. Next, a shared electronic whiteboard called ShareBoard is decribed. The goal of ShareBoard is to deliver a natural user interface to working with electronic whiteboards. Integrated within ShareBoard are input devices to recognise the movement of users in the surrounding space and for sensing 3D-gesture. ShareBoard can make use of media objects in NEMO. Last, we introduce the Modular Awareness Construction Kit. MACK is a framework for developing context aware, ambient intelligent systems that blend seamlessly with the users' everyday route, enabling unobtrusive in-situ interaction and facilitating enhanced cooperation and communication. In the future, MACK is to deliver contextual information to both NEMO and ShareBoard.

Introduction

In his seminal paper "The Computer for the 21st Century", Mark Weiser (1991) states that "the most profound technologies are those that disappear". He also introduces tabs, pads and boards as differently sized computing devices to explore this disappearing computer. He later clarifies that such devices only constitute "Phase 1" of ubiquitous computing, since it is "unlikely to achieve optimal invisibility" with such devices (Weiser, 1993).

Today, 20 years later, the promises of ubiquitous or ambient computing have yet to fully manifest themselves. Therefore, it is necessary to look at ambient interactive systems not only as invisible systems, but as part of an ecosystem comprised of both disappearing components and pads, tabs and boards in Weiser's sense, as well as more "conventional" computer systems.

This technical argument is supplemented by a user centered perspective (e.g. Bellotti & Edwards, 2001). Ubiquitous systems in the strictest sense pose new human-computer interaction problems. For example, how can explanations about the system state be delivered to the user if the only available interfaces are embedded into everyday objects like lighting fixtures or other home automation actuators?

In our research, we have therefore explored a design space ranging from disappearing, ambient computing over mobile devices to stationary devices such as interactive boards or presentation systems.

NEMO

The Network Environment for Multimedia Objects NEMO (Lob et al., 2010) is a smart media environment for semantically rich, contextualized, and device-specific access to and interaction with multimedia objects. NEMO is a distributed system that gives users access to media objects like images, videos, audio or text on a wide variety of different devices and in a wide variety of different contexts. Single media objects are called NEMO Information Objects (NIO). NIOs can be bundled in collections called NEMO Multimedia Objects or NMO.

On a very basic level, semantic information is stored implicitly through the organization of multimedia objects in NMOs. In addition, metadata like author name, license information, or free tags can be stored. The metadata can also cover the context of use. For example, a NMO used in a particular meeting situation could be annotated with details of the meeting (time, location, users accessing the NMO). For an even richer representation, the collections can contain links to concepts in an ontology. We have chosen to cater for linking NMOs or NIOs with linked data objects (Heath & Bizer, 2011). The canonical solution here is to link NMOs and NIOs to dbpedia[1].

Access to NIOs can be handled in a device-specific manner. To this end, information about device capabilities (like the types of media that can be played back, screen size, etc.) can be stored inside NEMO. The goal is to provide means for automatic manipulation of objects to fit them to the device requesting the object. Videos can be transcoded to formats supported; images can be scaled according to screen size. In the future, NEMO will also support objects with different representations. Devices with 3D-displays can then, for example, show a 3D-model of an object, whereas devices capable of 2D-output only would show a 2D-rendering of the same object.

ShareBoard

ShareBoard is an electronic whiteboard for local and remote collaborative work. It supports both pen- or touch-based interaction and interaction with mouse and keyboard as well as any combination thereof. Experimental support for 3D-gesture recognition is also available.

A primary design goal for ShareBoard is to provide a “clean” user interface, presenting the user with a prominent working area with almost no menus shown. Important functions can be called using a gesture to open up a pie menu anywhere on the surface. Other interaction elements, e.g. to add images to the drawing area, are provided in drawers on the sides of the working area. ShareBoard can optionally be coupled with NEMO. Images can be downloaded from NEMO and placed on the drawing surface. Vice versa, images

1 http://dbpedia.org/

from local storage that are added to the drawing surface are automatically added to NEMO and annotated with metadata of the current meeting context. Support for other media types is under implementation.

MACK

The Modular Awareness Construction Kit MACK (Schmitt et al., 2011) is a framework for developing context aware, ambient intelligent systems that blend seamlessly with the users' everyday routine, enabling unobtrusive in-situ interaction and facilitating enhanced cooperation and communication. MACK features a central server, the so-called awareness hub. Sensors and actuators (i.e. devices that change the environment or present information) are connected to this central hub.

On the awareness hub, different reasoners can access the sensor data and process it. Their findings are then used as input for higher level reasoners or to drive actuators. MACK makes no assumptions on the inner working of the different reasoners other than that they have to express their input and output in terms of a common ontology. The advantage of this process is that MACK can make use of different reasoning paradigms. In one example application, both rule-based and case-based reasoners are employed.

While MACK is not yet integrated with either NEMO or ShareBoard, the goal is to use MACK to deliver contextual information to other applications. For example, a meeting room can be automatically opened in ShareBoard and media downloaded from NEMO if a meeting on a particular topic is scheduled and the required participants are present.

Conclusions

The frameworks described are used to build various research prototypes. Applications range from shared whiteboards over presentation systems coupling stationary and mobile devices to ambient awareness systems. While

first user studies are promising, MACK has to be integrated in more applications to explore the interplay of disappearing and visible systems.

References

Bellotti, V. & Edwards, K. (2001). Intelligibility and Accountability: Human Considerations in Context-Aware Systems. Human–Computer Interaction Vol. 16, Nr. 2-4, pp. 193-212.

Heath, T. & Bizer, C. (2011) Linked Data: Evolving the Web into a Global Data Space. San Rafael, CA, USA: Morgan & Claypool.

Lob S., Cassens J., Herczeg M., Stoddart J. (2010). NEMO – The Network Environment for Multimedia Objects. In: Proceedings of IITM 2010, IIIT Allahabad, India. ACM, pp 245-249.

Schmitt F., Cassens J., Kindsmüller M. C., Herczeg M. (2011). Mental Models of Ambient Systems: A Modular Research Framework. In: Proceedings of CONTEXT 2011. Heidelberg: Springer, pp 278-291.

Weiser, M. (1991). The Computer for the 21st Century. Scientific American, September 1991, pp. 94-104.

Weiser (1993). Some Computer Science Issues in Ubiquitous Computing. Communications of the ACM Vol. 36, Nr. 7, pp. 75-84.

Erfolgsfaktoren mobiler Anwendungen für klein- und mittelständische Unternehmen

Eine quantitative Untersuchung

Lars Wendt

Universität Hildesheim, wendtl@uni-hildesheim.de
Am Deichhof 26
30459 Hannover
Lars_wendt@gmx.de

Zusammenfassung

Der Vortrag „Erfolgsfaktoren mobiler Anwendungen für klein- und mittelständische Unternehmen“ befasst sich mit der quantitativen Untersuchung von Erfolgsfaktoren für die Entwicklung von Apps oder mobiler Websites. In zuvor durchgeführten Experteninterviews wurden 45 Erfolgsfaktoren analysiert. Diese wurden unter Zuhilfenahme einer Online-Umfrage von Experten bewertet. Mit den Ergebnissen konnten nach Relevanz sortierte Rangfolgen der Erfolgsfaktoren erstellt werden.

Abstract

The presentation focuses on the quantitative analysis of success factors of mobile applications and websites for small and medium-sized enterprises. In previously conducted expert interviews 45 success factors were analyzed. The success factors were evaluated by experts with the aid of an online survey. Based on the results a ranking system sorted by relevance was created.

Einleitung

Der Zugriff auf Webinhalte über mobile Endgeräte steigt immer weiter an. Es wird prognostiziert, dass im Jahr 2013 ungefähr 75 Prozent der Weltbevölkerung auf das mobilen Web und nur noch 40 Prozent auf das Desktop-Internet zugreifen (Vgl. Fling 2009, 30). Im Jahr 2015 sollen außerdem eine Milliarde Menschen zu der Gruppe „Mobile-only-User“ gehören (Vgl. Mayer 2012, 9). Diese Zahlen verdeutlichen, dass es auch für klein- und mittelständische Unternehmen wichtig ist, Kunden eine mobil angepasste Website oder eine native App zur Verfügung zu stellen. Die vorgestellte Untersuchung befasst sich mit der Analyse von Erfolgsfaktoren für die Entwicklung von mobilen Anwendungen. Eine Erfolgsfaktorenanalyse ist notwendig um einen langfristigen Unternehmenserfolg sicherzustellen (Vgl. Heinrich und Stelzer 2009, 343). Grundlage dafür sind in der Regel Erfolgsfaktoren aus bereits durchgeführten Projekten (a.a.O., 349), welche in der vorgestellten Analyse erhoben wurden.

Methodik

Um eine Rangfolge von Erfolgsfaktoren erstellen zu können, wurden zwei Forschungsmethoden angewendet. Damit allgemein wichtige Erfolgsfaktoren eingegrenzt werden konnten, wurden zunächst Experten befragt. Die Relevanz, der aus den Interviews gewonnen Erkenntnisse, sollte dann in einer Online-Umfrage quantitativ untersucht werden.

Experteninterviews

Insgesamt wurden sieben Experten befragt. Die Grundlage für den Interviewleitfaden waren die vier Schlüsselbereiche für den IT-Erfolg nach Rockart (1982, 8 ff.): Service, Kommunikation, Personal und Positionierung. Darüber hinaus wurde weitere Literatur bezüglich Mobile Development mit einbezogen. Der Leitfaden sollte mögliche Schwerpunkte für mobile Entwicklungen abdecken. Das Ziel war es jedoch, eine offene Befragung durchzuführen, wodurch es ermöglicht werden sollte, dass neue Aspekte genannt werden.

Es wurden Informationen über den Tätigkeitsbereich, die Entwicklungen, Anzahl der Mitarbeiter, Größe des Unternehmens, Dauer der Projekte usw. erfasst. Zu den Teilnehmern gehörten Experten, welche an der Entwicklung von Apps als auch von mobilen Websites beteiligt waren und somit unterschiedliche Erfahrungen äußern konnten.

Als Ergebnis wurden 45 Erfolgsfaktoren analysiert, welche zum einen allgemein auf mobile Entwicklungen zutreffen. Einige sind jedoch nur bei bestimmten Anwendungsfällen, wie z.B. bei der Entwicklung unternehmensinterner Applikationen, relevant. Weiterhin widersprechen sich manche der erhobenen Erfolgsfaktoren. Die Gründe dafür sind die verschiedenen Disziplinen und Projekte der Teilnehmer.

Online-Umfrage

Wie die Ergebnisse der Experteninterviews verdeutlichten, sind nicht alle Erfolgsfaktoren für jedes mobile Projekt von Bedeutung. Die Erfolgsfaktoren sollten somit zusätzlich quantitativ überprüft werden. Um eine hohe Anzahl von Teilnehmern zu erreichen, wurde eine Online-Umfrage durchgeführt. Ziel der Umfrage war außerdem nach Relevanz sortierte Rangfolgen der Erfolgsfaktoren zu erstellen.

Zunächst wurden auch hier nach den wichtigsten Informationen über Teilnehmer und deren Entwicklungen gefragt. Dazu gehörten Fragen über Tätigkeiten, aktuelle Projekte, Unternehmensbranche sowie deren Größe. Ebenso wurde gefragt, ob die Teilnehmer selbst zu dem Entwicklungsteam gehörten oder Auftraggeber waren.

Die Erfolgsfaktoren wurden in sieben Fragengruppen unterteilt: „Projektplanung“, „Personal“, „Service“, „Entwicklung“, „native Applikationen“, „unternehmensinterne Applikationen“ und „Usability-Tests“. Um sehr komplexe Erfolgsfaktoren formulieren zu können, wurden Statements verwendet. Einige Erfolgsfaktoren mussten aufgeteilt werden, weshalb sich die Anzahl von 45 auf 50 erhöhte. Experten mussten auf einer fünfstufigen Likert-Skala ihre Zustimmung von „stimme überhaupt nicht zu“ bis „stimme voll zu“ angeben. Eine neutrale Bewertung „weder noch“ konnte auch gewählt werden. In Abhängigkeit von den angegebenen Erfahrungen der Teilnehmer wurden die Fragengruppen „native Applikationen“ und „unternehmensinterne Applikationen“ angezeigt oder ausgelassen.

Ergebnisse

Insgesamt haben 124 Personen an der Online-Umfrage teilgenommen. Es konnten jedoch nur 49 Umfragen ausgewertet werden, da die übrigen Fragebögen nicht vollständig oder akkurat beantwortet wurden. Auf Basis der erhobenen Daten wurden sieben Rangfolgen mit 37 Erfolgsfaktoren erstellt.

Einschätzung der Experten

Um die Wichtigkeit einer Erfolgsfaktorenanalyse zu verdeutlichen, sollen einige Ergebnisse kurz vorgestellt werden. Es ist interessant, dass die Relevanz des Erfolgsfaktors „Nutzerfeedback“ bei der Entwicklung höher eingeschätzt wird, als Kenntnisse über plattformspezifische Programmiersprachen. Die Differenz ist nicht erheblich. Die Bewertung zeigt außerdem, dass eine agile Entwicklung bei mobilen Projekten bevorzugt wird. Das Feedback durch Rezensionen, Kommentaren usw. wird stärker in die Entwicklung miteinbezogen, was erklärt, warum der Erfolgsfaktor „Konzeption“ erst an der dritten Stelle in dieser Rangfolge erscheint. Das Ergebnis zeigt, was auch in den Experteninterviews betont wurde: Eine mobile Anwendung wird viel schneller veröffentlicht und erst im Nachhinein verbessert.

„User Experience“ steht an erster Stelle in der Rangfolge „Service“. Damit eine mobile Website oder eine native Applikation genutzt wird, muss ein Nutzungserlebnis geschaffen werden. Die „Reduktion der Informationen und Inhalte“ wird als weniger relevant eingeschätzt. Der Erfolgsfaktor „Offline-Nutzung“ wurde als irrelevant eingeschätzt und ist somit nicht in die Rangfolge aufgenommen worden. Viel wichtiger war den Experten, dass aktuelle Informationen vorhanden sind.

Im Bereich „Usability-Tests“ wurden „Low Fidelity Prototypen“ am relevantesten eingestuft. Auch „Mehrkosten für Nutzertests“ sollten eingeplant werden. „Nutzertests unter realen Bedingungen“ wie Sonneneinstrahlung, Umgebungsgeräuschen usw. werden dagegen als eher unwichtig angesehen.

Zusammenfassung

Die Ergebnisse zeigen schließlich, dass Erfolgsfaktoren für die Entwicklung von mobilen Anwendungen nicht nur unterschiedlich wichtig sein können, sondern zum Teil nur auf bestimmte Anwendungsfälle zutreffen. Soll zum Beispiel eine unternehmensinterne Applikation entwickelt werden, müssen zusätzliche Erfolgsfaktoren berücksichtigt werden. Auch bei der Entscheidung für eine mobile Website oder native Applikation unterscheiden sich einige Erfolgsfaktoren. Je nach Anwendungsfall kann somit aus den Rangfolgen ausgewählt werden. Mithilfe der quantitativen Untersuchung können Unternehmen und Entwickler von den Experteneinschätzungen profitieren. Die Sortierung nach Relevanz vereinfacht die Auswahl der wichtigsten Erfolgsfaktoren.

Literaturverzeichnis

Fling, B. (2009). Mobile Design and Development. Practical Techniques for Creating Mobile Sites and Web Apps: Sebastopol: O'Reilly Media.

Heinrich, L. J., Stelzer, D. (2009). Informationsmanagement. Grundlagen. Aufgaben. Methoden. München: Oldenburg Wissenschaftsverlag.

Mayer, A. (2012). App-Economy. Millarden-Markt Mobile Business: München: Münchner Verlagsgruppe.

Rockart, J. F. (1982). The Changing Role of the Information Systems Executive. A Critical Success Factors Perspective. In: Sloan Management Review. S. 3-13.

Wendt, L. (2012). Erfolgsfaktoren mobiler Anwendungen für klein- und mittelständische Unternehmen. Eine quantitative Untersuchung. Hildesheim: Universität Hildesheim.

Gestaltungsrichtlinien für die Suche auf mobilen Webseiten:

Methodik und Ergebnisse einer Usability-Evaluation

Sören Dreyer

Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
SoerenDreyer@web.de

Zusammenfassung

Was sind die gestalterischen Herausforderungen, um den Nutzer optimal bei der Suche auf mobilen Webseiten im Smartphonebereich zu unterstützen? Dieser Beitrag liefert einen Überblick der wesentlichen Usability-Probleme in diesem Kontext, welche durch Usability-Tests mit der Restrospective-Thinking-Aloud-Methode analysiert und verifiziert wurden. Als Ergebnis werden die wichtigsten Erkenntnisse aus den empirischen Erhebungen erläutert und partiell mit bestehenden Konventionen des stationären Desktopbereichs kontrastiert.

Abstract

Which creative challenges does one face when it comes to aiding users optimally during their search on mobile websites with smartphones? The paper gives an overview of elementary usability-issues in this context, for which usability-tests were analyzed and verified using the Retrospective-Thinking-Aloud-Method. Results include illustrating the most important findings, based on empirical surveys, which will partially be contrasted with established conventions in the area of stationary desktop pcs.

Einleitung

Die Zahl der Smartphonenutzer weltweit steigt stetig an und somit auch die Zahl derjenigen Nutzer, die ihre Informationsbedürfnisse auf mobilen Endgeräten stillen möchten. Webtrekk GmbH, ein Anbieter für High-End-Webanalysen, bestätigte im 3. Quartal 2012 für Deutschland, dass bereits mehr als jeder zehnte Webzugriff über mobile Geräte stattfindet (11,4%) (vgl. Oschatz, 2012). Egal ob mit Apples Betriebssystem iOS, Googles Android, Windows Phone oder sonstigen mobilen Betriebsplattformen, ein wichtiger Bestandteil des Webs, mobil oder stationär, bildet die Suche nach Informationen. Die meisten Usability-Probleme bei der Nutzung des Webs liegen im Bereich der Suche und der allgemeinen Informationsarchitektur (vgl. Nielsen & Loranger, 2006: 127f). Daran zeigt sich, dass großer Handlungsbedarf bei der Optimierung dieses Bereichs besteht.

Ein Mensch verfügt über erworbene mentale Modelle bereits bestehender Systeme und stellt Analogien zu anderen, ähnlich bedienbaren Systemen her (vgl. Carroll & Reitman, 1987: xif). Daher ist es nur logisch, dass bereits bestehende Suchmechanismen, die im stationären Desktopbereich zu finden sind, ebenso im mobilen Bereich angewendet werden sollten. Dies sollte jedoch unter bestimmten Voraussetzungen geschehen, da Smartphones, um die es in diesem Beitrag ausschließlich gehen soll, in ihrer gleichzeitigen Darstellungsvielfalt aufgrund der kleinen Bildschirmgröße maßgeblich beschränkt sind. So sollten mobile Webseiten eben für jene Endgeräte optimiert sein. Dies bedeutet, dass bestimmte Features für den mobilen Bereich entfernt werden müssen, welche nicht ausschlaggebend für die mobile Nutzung sind. So sollten Inhalte auf das Allernötigste reduziert werden und zusätzliche Informationen durch Verlinkung auf *zusätzlichen* Seiten der Webseite platziert werden. Außerdem ist es wichtig, Bedienelemente wie Buttons stets groß zu halten, um dem sogenannten 'Fat Finger' Problem vorzubeugen, welches impliziert, dass Elemente stets eindeutig anwählbar sein sollten, ohne unbeabsichtigt andere Elemente einer mobilen Webseite zu berühren (vgl. Nielsen & Budiu, 2013: 20). Die Wichtigkeit der Optimierung verdeutlicht sich insbesondere im Bereich der Mobile Commerce Webseiten wie z.B. Amazon.de oder Zalando.de. Laut Nielsen & Loranger (2006) zeigt sich,

dass die 'Conversion Rate'[1] im Bereich der Smartphones bei nur 1,4% liegt. Im Vergleich liegt die Rate im stationärem Desktopbereich bei 3,5% (Tablets 3,2%). Das liegt zum einen daran, dass Smartphones überwiegend für oberflächlichere Aufgaben wie Spiele oder Soziale Netze verwendet werden, zum anderen aber auch am geringeren 'ease of use' im mobilen Bereich (vgl. Nielsen & Budiu, 2013: viiif). In Bezug auf die 'Conversion Rate' ist die Informationssuche ein wichtiger Bestandteil, der unabdingbar ist, erfolgreich einen Kauf zu tätigen, denn: „If users can’t find a product on a mobile site, they assume the company doesn’t sell it and go elsewhere.“ (Nielsen & Budiu, 2013: 20). Der vorliegende Artikel behandelt zentrale Aspekte, die in Dreyer (2013) ausführlich dargestellt werden.

Methodik

Im Folgenden wird die empirische Vorgehensweise erläutert, die zum Erkennen von Usability-Problemen eingesetzt wurde. Die Methodik gliedert sich in einen explorativen Test, einen Pilottest, einer Vorbefragung, dem eigentlichen Usability-Test mithilfe der Retrospective-Thinking-Aloud-Methode (RTA) und einem User-Experience-Questionnaire (UEQ). Die Auswahl dieser Methoden liegt zum einen darin begründet, den methodischen Ablauf in Bezug auf die Fragestellungen und des technischen Aufbaus zu prüfen und zu optimieren (explorativer Test und Pilottest). Zum anderen ermöglicht sie eine gute Kombination aus quantitativen und qualitativen Daten, die gemeinsam ein adäquates Gesamtbild, der in dem getesteten Kontext aufgetretenen Usability-Probleme, in Bezug auf Erfahrung der Nutzer liefern (Vorbefragung und Usability-Test). Da die User Experience (UX) ein Bestandteil der Usability ist, wird der UEQ als quantitativ, subjektiver Gesamteindruck des Nutzungserlebnisses hinzugezogen und ebenso mit den bereits gesammelten Daten in Relation gesetzt (vgl. Nielsen, 2012).

---

1 Die 'Conversion Rate' ist eine Kennzahl in Prozent von Nutzern, die eine gewünschte Aktion getätigt haben. Beispielsweise repräsentiert sie im Mobile Commerce diejenigen Nutzer, die einen Kauf abschließen (vgl. Nielsen & Budiu, 2013: viii).

Explorativer Test

Als Testobjekt für den Usability-Test wurden vier Webseiten, unter anderem aufgrund vorangegangener Abschlussarbeiten im Bereich von Gestaltungsrichtlinien im mobilen Kontext (vgl. Jaspers, 2012) ausgewählt. Davon gliedern sich zwei in die Kategorie der Mobile Commerce Seiten (Amazon.de und Zalando.de) und zwei in die Kategorie der Informationsseiten (Chefkoch.de und Kochbar.de). Zu Beginn wurden pro Seite vier Testszenarien (Aufgaben) konzipiert, die später gemeinsam mit den Testpersonen geprüft werden sollten. Ein explorativer Pre-Test dient zur Evaluierung und zum Messen der Effektivität der Testszenarien und ggf. dazu, Probleme im Voraus zu erkennen und neue Szenarien und/oder Fragestellungen zu entwickeln, falls Situationen und/oder Probleme auftreten, an die vorher nicht gedacht wurden und die später in den eigentlichen Usability-Test mit einfließen können (vgl. Rubin, 1994: 31f). Es wurden hierfür drei Testpersonen unter nahezu realen Testbedingungen getestet (siehe *Usability-Test mithilfe der RTA-Methode*). Der einzige Unterschied zum eigentlichen Usability-Test bestand darin, dass keine Audio- und Videoaufzeichnung erfolgte und die Testperson während des 'Durchspielens' der Szenarien laut dachte, während der Testleiter sich Notizen machte. Generell fand der explorative Test nicht unter so stark kontrollierten Bedingungen statt wie der Haupttest mit der RTA-Methode. Durch den explorativen Test wurden somit folgende Fragestellungen entwickelt und verifiziert:

- Wie gut unterstützt die mobile Suche den Nutzer generell (Auto-Suggest, Autocomplete etc.)?
- Welche Probleme treten bei der Suche im Allgemeinen und im Speziellen bei den Search Engine Result Pages (SERPs) auf?
- Sind die Terminologien für die Nutzer verständlich?
- Wird der Nutzer bei den SERPs in Bezug auf die Gestaltung (Durchblättern der Ergebnisse, Bilder, Layout etc.) unterstützt?
- Wann wird die manuelle Suche, wann die Kategorie-Suche genutzt?
- Wann / Wie / Welche Probleme treten bei parametrischer Suche auf (Probleme der Facettennavigation)?
- Wann / Wie / Welche Probleme treten bei Sortieren- vs. Filterfunktion im Einzelnen und in Kombination auf?
- Wie gut funktioniert der iterative Suchprozess?

- Was passiert, wenn nichts gefunden wird in Bezug auf die Reaktion des Nutzers und die Unterstützung der Webseite hinsichtlich der Anfrageformulierung, Ergebnisverfeinerung (Filtern, Sortieren), alternativer Ergebnisse und allgemeiner Verständlichkeit sowie Darstellung.

Pilottest

Anhand der Fragestellungen wurden die Testszenarien optimiert und schließlich im Pilottest das erste Mal unter Realbedingungen getestet. Der Pilottest dient zur Überprüfung des technischen Equipments und zum Aufspüren etwaiger Fehler im gesamten Ablauf des Usability-Tests. Die Erfahrung der Testpersonen für den Pilottest spielt eine unwesentliche Rolle. Des Weiteren ist es eine gute Übung für den Testleiter, den gesamten Test einmal durchzuführen (vgl. Rubin, 1994: 225).

Vorbefragung

Der erste Teil des Usability-Tests besteht aus einer Vorbefragung, um einige demographische Daten und Fakten über die jeweilige Testperson (TP) zu bekommen, damit sie später als Anfänger oder Experte in Bezug auf die Nutzung mobiler Endgeräte eingestuft werden kann. Diese Einteilung erfolgt aufgrund der mitgeteilten Nutzungsdauer, -orte und -arten der von den Testpersonen privat genutzten Smartphones, sowie der Frage, ob die TP bereits mobil einen Kauf tätigte und durch die Anzahl installierter Applikationen.

Usability-Test mithilfe der RTA-Methode

Im Anschluss der Vorbefragung beginnt der Hauptteil des Usability-Tests. Getestet wurde mit insgesamt 14 Probanden. Die TP bekommt nacheinander zwei Webseiten gezeigt, auf denen sie jeweils vier Aufgabenszenarien durchspielen soll. Nach jeder Aufgabe soll die TP subjektiv einschätzen, wie schwer oder wie leicht ihr die Aufgabe fiel. Die technische Testumgebung besteht aus einem iPhone4S mit iOS 6.0.1, Mikrofon, Webcam und einem Laptop, mit deren Hilfe der Testleiter mit der Aufnahmesoftware 'Morae' von Techsmith[2] sowohl die Aktionen als auch die TP selbst sehen und aufnehmen kann. Es wird sowohl der Screen des iPhone4S als auch die TP durch eine Webcam in Bild und Ton aufgenommen, um hinterher bei der Auswertung Gestiken, Mimiken sowie etwaige Kommentare zu erkennen und interpretie-

2 http://www.techsmith.de/morae.html

ren zu können. Durch die RTA-Methode bewältigt die TP alle vier Aufgaben pro Webseite zunächst ohne jeglichen Kommentar, d.h. sie denkt währenddessen nicht laut. Die Aufnahmen werden erst nach Bewältigung der Aufgabenszenarien zusammen mit dem Testleiter angeschaut, um anhand des Videomaterials Probleme und Fragen abzuarbeiten. Dies hat den Vorteil gegenüber der CTA-Methode (Concurrent-Thinking-Aloud), bei welcher die TP während der Bewältigung der Aufgabe laut ihre Gedanken äußern muss, dass die TP während des Tests nicht durch ihre eigenen Worte und verbalisierten Gedanken abgelenkt ist, sondern sich voll und ganz auf die Natürlichkeit der Testumgebung konzentrieren kann. Ein weiterer Vorteil der RTA-Methode ist, dass die Aufzeichnungen des Probanden und seine Aktionen hinterher zusammen mit dem Testleiter angeschaut werden und somit gezielt Fragen zu einzelnen Aktionen und Problemen gestellt werden können, ohne den Probanden in seinen natürlichen Aktionen zu beeinflussen. Zusätzlich kommt die Möglichkeit hinzu, jederzeit die Aufnahme zu pausieren oder zu einem früheren/späteren Zeitpunkt der Aufnahme zu wechseln. Der einzige nennenswerte Nachteil der RTA-Methode ist, dass der gesamte Usability-Test mehr Zeit benötigt, da quasi zwei Durchgänge (Durchspielen der Szenarien ohne jeglichen Kommentar und die Aufnahmen zusammen mit Testleiter anschauen und kommentieren) vollzogen werden müssen (vgl. Zühlke, 2004: 130; vgl. Hartson & Pyla, 2012: 442ff). Beim zweiten Durchgang werden wiederum der Screen des Laptops, auf dem nun das zuvor aufgenommene Video und der Screen-Film des iPhone4s mit den vorangegangenen Aktionen der TP zu sehen sind und die nun neuen Kommentare des Probanden und Fragen des Testleiters parallel zu dem Gezeigten aufgezeichnet. Diese Aufzeichnung wurde mithilfe der Software Screencastomatic[3] realisiert. Der Wechsel vom ersten zum zweiten Durchgang der Aufnahmeprozedur musste während des Tests relativ zügig geschehen, da die TP während der technischen Umstellung warten muss. Aufgrund der RTA-Methode ist es dem Testleiter möglich, hinterher die Aufnahme in Ruhe auszuwerten. Ferner wurden zu jeder Aufgabe noch quantitative Daten ausgewertet, wie etwa die Zeit, die eine TP pro Aufgabe brauchte.

3 http://www.screencast-o-matic.com

User-Experience-Questionnaire (UEQ)
Die Testpersonen sollten nach Absolvieren der Aufgaben jeder getesteten Webseite den Gesamteindruck des Nutzungserlebnisses mithilfe des UEQ[4] bewerten. Diese quantitative Abschlussmessung wird im Allgemeinen nach Abschluss der Testszenarien eines Usability-Tests empfohlen, damit die Testpersonen einen Gesamteindruck, in diesem Falle über die UX, abgeben können (vgl. Tullis & Albert, 2008: 135f). Dies erscheint insbesondere deswegen sinnvoll, da laut Nielsen & Loranger (2006) die Suche ein wichtiger Teil der Web-User-Experience ist.

Auswertung
Bei der qualitativen Auswertung der Testergebnisse wurden alle aufgenommenen Videos nacheinander angeschaut und im Einzelnen der Benutzerweg und auftretende Probleme/TP/Szenarien in einer Excel-Tabelle dokumentiert. Hinterher wurden ähnliche Probleme zunächst oberflächlich gruppiert und anschließend einzelnen Usability-Problemen zugeordnet (Häufigkeit von Problemen). Ferner wurde Aufgabenszenarien besondere Aufmerksamkeit geschenkt, wenn Testpersonen diese abbrachen oder wenn eine Aufgabe vom Probanden nur unter offensichtlichen Schwierigkeiten oder mithilfe des Testleiters durchgeführt werden konnten (Auswirkung und Beständigkeit des Problems) (vgl. Nielsen & Loranger, 2006: 123). Die analysierten Probleme, die auf den vier ausgewählten Webseiten mithilfe der hier vorgestellten Methode analysiert werden konnten, wurden mit bereits bekannten Usability-Problemen und Gestaltungrichtlinien im mobilen und stationären Bereich aus der Literatur verglichen und Rückschlüsse gezogen, welche Richtlinien verifiziert, geändert oder neu entwickelt werden müssen, um eine bestmögliche Suche auf mobilen Webseiten zu gewährleisten.

Ergebnisse

Nach Auswertung der empirischen Vorgehensweise respektive dem Usability-Test sind folgende Ergebnisse im Kern festzuhalten:

---

4 http://www.ueq-online.org

Acht Testpersonen wurden aufgrund der Vorbefragung als Experten, sechs als Anfänger klassifiziert (siehe *Vorbefragung*). Das Durchschnittsalter der TP betrug 25,5 Jahre. 43% sind weiblich, 57% männlich. Bereits mobil eingekauft hatten bereits 42,9% aller Befragten. Im Folgenden werden nun die drei am häufigsten Usability-Probleme erläutert.

Die häufigsten Probleme der getesteten Webseiten (Amazon.de, Zalando.de, Kochbar.de, Chefkoch.de) mit zirka 27% aller in diesem Test gefundenen Usability-Probleme wurden bei der Ergebnisdarstellung und der Zufriedenheit der Ergebnisse allgemein in Bezug auf Verständlichkeit, Initialsortierung der Ergebnisse, der Terminologie, der Ergebnismenge und die Unterstützung des iterativen Suchprozesses allgemein festgestellt. So finden Nutzer der mobilen Seite Chefkoch.de, laut gestelltem Aufgabenszenario, keine Rezepte bei Eingabe des Wortes 'Sudan', wohl jedoch drei Rezepte nach Eingabe des Suchwortes 'Sudanesisch'. Das Problem wurde im Test als besonders schwerwiegend klassifiziert, da es in 50% der Fälle zum Abbruch des Aufgabenszenarios führte, weil die Testpersonen keine Möglichkeit fanden, die Aufgabe zu lösen. Auffällig ist, dass es sich hierbei zunächst nicht um ein spezifisches Problem bei mobilen Endgeräten handelt, sondern um ein generelles Problem des Stemming-Verfahren5. Im mobilen Kontext betrachtet wird der Nutzer in diesem Beispiel jedoch nicht ausreichend durch einen iterativen Suchprozess unterstützt, nachdem seine Suche erfolglos blieb. Die Suche sollte zwar vom Desktopbereich hin zu kleineren Monitoren reduziert und optimiert werden, nicht jedoch im Umfang wichtiger Funktionen. Gestaltungsrichtlinie hierfür: Zeige klar und deutlich, dass keine Ergebnisse gefunden wurden und biete dem Nutzer durch Interaktionsmöglichkeiten an, weitere Ergebnisse zu bekommen. Jene Richtlinie wurde bereits von Jaspers (2012) aufgestellt. Des Weiteren ist es ratsam, den Suchalgorithmus bezüglich Teilwörter einer Suchanfrage zu optimieren. Diese Richtlinie ist ebenso im stationären Desktopbereich ratsam. Im mobilen Kontext sollten weitere Ergebnisvorschläge, falls nichts gefunden wird, durch Reduzierung von Wörtern und Darstellung von Bildern angepasst werden (vgl. Nudelman, 2011: 5).

---

5 Das Stemming-Verfahren reduziert Varianten eines Wortes auf dessen Wortstamm. Hier: Sudan ist im Wort Sudanesisch inkludiert und sollte zu gleichen Suchergebnissen führen (vgl. Kowalski, 2011: 76f).

Die zweithäufigsten Probleme mit zirka 20% aller in diesem Test gefundenen Usability-Probleme wurden bei der Kategorie-Suche und der Filtern- / Sortierfunktion festgestellt, also bei denjenigen Funktionen, die dem Nutzer helfen sollen, eine Einschränkung ihrer Suchanfrage und Ergebnisse zu gewährleisten, um bestimmte Produkte bzw. Informationen auf einer Webseite zu finden. Im Detail beziehen sich diese Probleme auf die Handhabung, das Fehlen von Filterarten oder dem gänzlichem Fehlen der Filterfunktion und/oder von Kategorien. Laut Nudelman (2011) ist gerade das Phänomen des „Filtering by Sorting“ bei den Nutzern ein vielfach beobachtetes. Der Nutzer macht sich keine komplexen Gedanken über den Unterschied von Filtern und Sortieren, jedoch tut dies ein Großteil der Webseitenentwickler, was zur Folge hat, dass viele Webseiten, wie z.B. Amazon.de, die Funktion von Sortieren und Filtern getrennt betrachten und eben auch getrennt gestalten, was unweigerlich zu Usability-Problemen führt. Der Nutzer filtert jedoch (auch) durch Sortieren, indem er zum Beispiel nach Preis sortiert und somit für ihn die Produkte mit den niedrigsten Preisen 'gefiltert' erscheinen. Im Test stießen viele Testpersonen auf das Problem, dass es auf der mobilen Webseite von Amazon.de erst möglich war zu sortieren, nachdem eine Kategorie ausgewählt wurde. Dies zeigt auch, warum die Kategorie-Suche so eng verzahnt mit der Filtern- und Sortierfunktion ist und stets im Verbund betrachtet werden sollte (vgl. Nudelman, 2011: 155f). Somit lässt sich die Gestaltungsrichtlinie verifizieren, dass die Sortierfunktion stets präsent und in Abhängigkeit der Filterfunktion betrachtet werden sollte.

Die dritthäufigsten Probleme mit zirka 17% aller in diesem Test gefundenen Usability-Probleme beziehen sich auf die Suchleiste und deren Funktion in Bezug auf Autocomplete/Autosuggest. Es wurde festgestellt, dass nur auf Zalando.de eine Autosuggest-Funktion implementiert war, was gerade in Bezug auf die Minimalisierung der Texteingabe im mobilen Bereich eine sehr wichtige Funktion ist, da laut Nielsen (2013) langes Tippen auf dem Smartphone noch immer zu den meistgehassten Pflichten der Nutzer gehört, was im Usability-Test bestätigt werden konnte. Des Weiteren kommt hinzu, dass nicht immer ein vernünftiger Korrekturvorschlag angezeigt wurde, nachdem die Testperson sich vertippt hat. Zalando.de zeigte bei einem Aufgabenszenario leider ein Negativbeispiel, wie Autosuggest nicht ganz optimal eingesetzt wird. So sollte die TP nach Zubehörartikel für die Sportart 'Yoga' suchen. Durch Eingabe des Wortes 'Yoga' bekam der Nutzer allerdings keine Ergebnisse, sondern die Empfehlung zur Desktopversion der Webseite zu wechseln. Hätte der Nutzer jedoch den von Zalando.de angebo-

tenen Autosuggest angenommen (z.B. 'Yoga Sports'), so hätte er Ergebnisse bekommen. Hier scheint Zalando.de dem Nutzer seine Suchvorschläge regelrecht aufzuzwingen, was nicht im Sinne dieser Funktion sein sollte. Mittlerweile funktioniert die Suche bei Eingabe des Suchbegriffs 'Yoga' jedoch einwandfrei. Ein weiteres Problem war, dass die Suchleiste nach einer Suche nicht mehr sichtbar war und erst durch Klicken auf ein Lupensymbol zum Vorschein kam. 57% aller Probanden, welche die Suchleiste auf Zalando.de nutzten, ein Großteil davon die Anfänger, hatten damit im Test Schwierigkeiten und verwechselten die Suchmaske des Internetbrowsers mit der Suchmaske der mobilen Webseite. Somit bleibt festzuhalten, dass die Suchleiste stets gut sichtbar platziert und einsehbar sein sollte. Sie sollte sowohl im oberen Bereich der Webseite, als auch im unteren Bereich vorhanden sein.

Die restlichen Probleme beziehen sich auf den Bereich, dass Äquivalenzen zum Desktopbereich nicht gegeben waren, wie zum Beispiel eine einfache Sortierung nach Preis. Außerdem kam es vor, dass Testpersonen ungewollt von der mobilen Seite auf die Desktopseite umgeleitet wurden. Ferner gab es teilweise Probleme bezüglich der Klickbarkeit ('Fat-Finger' Problematik). In seltenen Fällen wurde eine Pinch-To-Zoom Funktion vermisst, was durch eine angemessene Schriftgröße und Darstellung der Buttons jedoch obsolet werden würde. Insgesamt wurden bei 20 von 112 geprüften Testszenarien schwerwiegende Usability-Probleme diagnostiziert, die zum Abbruch führten, was einem Anteil von 17,9% aller Aufgaben entspricht.

Fazit

Viele der herausgearbeiteten und hier im Auszug gezeigten Usability-Probleme hinterlassen den Eindruck, dass der Bereich der mobilen Webseiten tendenziell inkonsequent von den Entwicklern angegangen wird. Dies lässt vermuten, dass mobile Webseiten in der Entwicklung noch nicht hoch genug priorisiert werden, was allerdings fatal sein könnte, wenn man sich die rasante Verbreitung des mobilen Webs anschaut. Diese Vermutung wird insbesondere durch den relativ hohen Anteil (17,9%) der schwerwiegenden Usability-Probleme, die zum Abbruch von Aufgabenszenarien führten, untermauert. Die hier angewendete empirische Erhebungsmethode, insbesondere die RTA-Methode für Usability-Tests, wurde als durchweg positiv und

effektiv empfunden. Der gesamte Workflow der Tests lief mit dem in diesem Beitrag vorgestellten Equipment reibungslos ab und brachte brauchbare und aufschlussreiche Ergebnisse zu Tage. Die Möglichkeit, dass die Testpersonen durch die RTA-Methode ohne Beeinflussung des lauten Redens ihre Testaufgaben in Ruhe absolvierten und im Nachhinein ihre Gedanken anhand des Videos erläuterten, funktionierte in dem hier vorgestellten Testrahmen problemlos und lieferte zugleich realitätsnahe, authentische Nutzerdaten. Somit wurde sie hier als eine effiziente Erhebungsmethode wahrgenommen und machte aufgrund der natürlichen Testatmosphäre den Nachteil der längeren Testzeiten, zumindest im Rahmen kleinerer Testgruppen, mehr als wett (vgl. Hartson & Pyla, 2012: 441f).

References/Literaturverzeichnis

Carroll, J. M. /. Reitman Olsen, J. (1987): Mental Models in Human-Computer Interaction. Research Issues About What the User of Software Knows. Washington D.C.: National Academies Press.

Dreyer, S. (2013). Gestaltungsrichtlinien für die mobile Suche: Evaluation und Vergleich von alternativen Ansätzen. Mag.-Arb. Stiftung Universität Hildesheim. Institut für Informationswissenschaft und Sprachtechnologie (erscheint im 3.Quartal 2013).

Hartson, H. R. / Pyla, P. S. (2012). The UX Book. Process and guidelines for ensuring a quality user experience. Amsterdam, Boston: Elsevier.

Jaspers, T. (2012). Gestaltungsrichtlinien für mobiles Web: Evaluation von alternativen Navigationsansätzen. Mag.-Arb. Stiftung Universität Hildesheim. Institut für Informationswissenschaft und Sprachtechnologie.

Kowalski, G. (2011). Information retrieval architecture and algorithms. New York: Springer.

Nielsen, J. (2012). Usability 101: Introduction to Usability. http://www.nngroup.com/articles/usability-101-introduction-to-usability/, zuletzt geprüft am 13.04.2013.

Nielsen, J. / Budiu, R. (2013). Mobile usability. Berkeley, CA: New Riders.

Nielsen, J. / Loranger, H. (2006). Web Usability. München: Addison-Wesley.

Nudelman, G. (2011). Designing Search. UX Strategies for eCommerce Succes. New York: Wiley.

Oschatz. A. (2012). Kennzahlen zum Deutschen Mobile-Markt mit Charts. Hg. v. Webtrekk GmbH. http://mobilemetrics.de/2012/10/02/kennzahlen-zum-deutschen-mobile-markt-mit-charts/, zuletzt geprüft am 15.04.2013.

Rubin, J. (1994). Handbook of usability testing. How to plan, design, and conduct effective tests. New York: Wiley.

Tullis, T. / Albert, B. (2008). Measuring the user experience. Collecting, analyzing, and presenting usability metrics. Amsterdam, Boston: Elsevier.

Zühlke, D. (2004). Useware-Engineering für technische Systeme. Berlin: Springer.

Session 4:

Social Web

Aufbruch nach Zyren

Game-based Learning in der Hochschullehre

Lisa Orszullok, Kathrin Knautz, Simone Soubusta

Heinrich-Heine-Universität Düsseldorf
Universitätsstr. 1
40204 Düsseldorf
Lisa.Orszullok@hhu.de, Kathrin.Knautz@hhu.de, Simone.Soubusta@hhu.de

Zusammenfassung

In diesem Beitrag wird mit der Verwendung von Spielmechaniken im universitären Kontext (game-based Learning) ein innovatives Lehr-Lernprojekt vorgestellt, in welchem einer neuen Generation von Lernenden, den Digital Natives, Informationskompetenz als eine der Schlüsselqualifikationen des 21. Jahrhunderts vermittelt werden kann. Der kompetente Umgang mit der Ressource Wissen ist in der heutigen als Wissensgesellschaft bezeichneten Gesellschaftsform unabdingbar. Es liegt nahe, dieser neuen Generation von Lernenden mit einer innovativen Lehrmethode und den vorhandenen technologischen Möglichkeiten und Ressourcen entgegenzukommen, um das notwendige Wissen zu vermitteln. Einhergehend mit dem Projekt wurde eine umfassende Evaluation konzipiert, die die Besonderheiten des Konzepts ausführlich berücksichtigt.

Abstract

In this paper, we present an innovative didactic project based on the use of game mechanics in university context (game-based learning) in which a new generation of students, the digital natives, should learn information literacy as a key skill of the 21th century. The proficient handling of the resource knowledge is indispensable in our society – also known as knowledge society. It stands to reason that this new generation of learners should be accommodated with innovative teaching methods and the use of the existing technological possibilities and resources to impart required knowledge. At-

tended by the project there was a comprehensive evaluation draft with respect to the characteristics of the concept.

Leben in der Wissensgesellschaft

Die Ressource Wissen bildet die Basis für unser ökologisches und soziales System. Daraus folgt, dass wir uns in einem wissensbasierten Zeitalter befinden, dessen gesamte Gesellschaft darauf ausgerichtet ist, Informationsinfrastrukturen zu verstehen, aufzubauen und zu nutzen. Die kompetente Nutzung von Information- und Kommunikationstechnologien *(information and communication technology* – ICT) bildet hierbei einen wesentlichen Bestandteil der wirtschaftlichen und gesellschaftlichen Entwicklung. Für die Generation der Digital Natives (Prensky, 2001a), die mit diesen technischen Errungenschaften aufgewachsen ist, stellt dies meist kein Problem dar. Prensky (2001b) charakterisiert diese Generation wie folgt:

> „Digital Natives [are] accustomed to the twitch-speed, multitasking, random-access, graphics-first, active, connected, fun, fantasy, quick-payoff world of their video games, MTV, and Internet (…)".

Sie nehmen Smartphones, das Internet, Social Software etc. als einen alltäglichen Bestandteil ihres Lebens wahr und sind daran gewöhnt, jederzeit mit all ihren Freunden und Bekannten in Kontakt treten zu können und zu jedem Zeitpunkt und von überall Informationen zu erhalten (Gaston, 2006). Dies erfordert eine ausgeprägte Multitasking-Fähigkeit, um die verschiedenen Informationsflüsse bestehend aus privaten Informationen von Freunden und Bekannten innerhalb von sozialen Netzwerken und Fakteninformationen aus dem gesamten Internet zu beherrschen. Diese Fertigkeit haben sich viele junge Erwachsene dieser Generation nahezu perfekt angeeignet.

Um sich in dieser von komplexen Informationsinfrastrukturen geprägten Welt zurechtzufinden und somit innerhalb dieser interagieren zu können, sind aber nicht nur reine ICT-Kenntnisse notwendig. Die richtigen Informationen zu ermitteln, zu bewerten und anzuwenden, sowie wiederzufinden und neu zu strukturieren, sind dabei essentielle Fähigkeiten (Gust von Loh & Stock, 2013). Dadurch nimmt die Informationskompetenz eine zentrale Rolle ein. Damit einhergehend ist das lebenslange Lernen ein Kernelement der

Informationskompetenz (Heinze & Schnurr, 2010). Lernen ist somit als zentrale Fähigkeit in der heutigen Gesellschaft anzusehen, welche möglicherweise sogar noch bedeutsamer ist als je zuvor.
Mit den Digital Natives existiert in den Hochschulen eine Generation von Studierenden, die zwar eine hohe Affinität zur ICT-Infrastruktur aufweist, jedoch in anderen Bereichen wie der Informationskompetenz Schwächen aufweist. Ziel muss es sein, eine motivationsfördernde Lernumgebung zu schaffen, um Schlüsselqualifikationen adäquat vermitteln zu können.

Freude am Lernen - das Flow-Erleben

Berücksichtigt man die zuvor aufgezeigten Eigenschaften der Digital Natives im Vergleich zu den vorherigen Generationen, so lässt sich eine veränderte mediale Erfahrungswelt der Studierenden erkennen. Jedoch reicht dieses Wissen nicht allein aus, um die Motivation zurück in den Unterrichtsraum zu bekommen. Einerseits muss man Methoden entwickeln, die diesen Veränderungen entgegenkommen und andererseits muss man verstärkt berücksichtigen, was die Lernmotivation bedingt.

Ein entscheidender Aspekt für die Lernfreude stellt die Motivation dar. Hierbei wird zwischen der extrinsischen und intrinsischen Motivation unterschieden. Extrinsisch-motivierte Handlungen sind durch einen Reiz, einen Stimulus, bedingte Aktionen, die darauf ausgerichtet sind, eine bestimmte Konsequenz zu erzeugen (Schlag, 2009). Diese Konsequenzen sind meist das Empfangen positiver Bekräftigungen des zuvor ausgeführten Verhaltens, wohingegen allerdings auch das Vermeiden von negativen Folgen möglich ist. Demgegenüber steht die intrinsische Motivation, welche dazu führt, dass eine bestimmte Handlung um ihrer selbst willen ausgeführt wird, ohne dass ein Stimulus von außen sie anstößt. Intrinsische Handlungen werden mit einem gewissen Selbstverständnis durchgeführt, welches darauf beruht, dass bestimmte Bedürfnisse erfüllt werden, wie beispielsweise Spaß zu haben oder Herausforderungen zu bewältigen zu wollen (Myers, 2004). Eng mit der intrinsischen Motivation verbunden ist das sogenannte Flow-Erlebnis (Csikszentmihalyi, 1985). Das Flow-Erleben ist das vollständige Aufgehen in einer Tätigkeit. Entscheidend ist hierbei, dass man weder überfordert noch unterfordert wird, sondern die Tätigkeit als reines Vergnügen empfindet. Für den Lernprozess bedeutet das, dass man zu lernen beginnt und Wissen erwirbt, welches wiederum auf neue Aufgaben angewendet werden kann. Dabei steigt

mit dem erworbenen Kompetenzstand auch der Schwierigkeitsgrad der Tätigkeiten. Die Anforderungen einer Aufgabe müssen zu jedem Zeitpunkt in einem angemessenen Verhältnis zu den eigenen Fähigkeiten stehen. Unterstützend wirkt dabei, wenn der Lernende unmittelbare Rückmeldung erhält und im Zweifelsfall darauf hingewiesen wird, falls ein nicht lösungsführendes Verhalten vorliegt. Außerdem ist es erforderlich, die Tätigkeit auf ein bestimmtes Ziel auszurichten, auf welches man hinarbeitet.

Generell gilt, die Motivation nicht nur als fördernd beim Lernprozess anzusehen, sondern sie als einen zentralen Bestandteil dafür wahrzunehmen, dass sich Lernerfolge verzeichnen lassen. Mit entsprechenden Methoden und einer hohen Motivation werden darüber hinaus höhere Kompetenzstufen erworben (Artelt, Baumert, Julius-McElvany, & Peschar, 2003).

Neue Wege in der Hochschuldidaktik

Es stellt sich nun die Frage, wie man die Lernmotivation der Lernenden fördern und vor allem aufrechterhalten kann. Eine Antwort hierauf kann das game-based Learning sein. Dazu werden Mechanismen aus dem Gamedesign in die Lehre eingebunden. Werden diese Verfahren allgemein in spielfremden Kontexten eingesetzt, um somit das Engagement zu fördern, spricht man auch von Gamification (Deterding, Sicart, Nacke, O'Hara, & Dixon, 2011).

Dabei dienen das Spiel sowie das game-based Learning nicht nur einem reinen Unterhaltungszweck, sondern nehmen eine bestimmende Rolle in der menschlichen Entwicklung ein. Für den homo ludens („spielender Mensch") stellt das Spielen einen zentralen Aspekt im kulturellen Leben dar, da schon seit Tausenden von Jahren das Spiel zum Erwerb von überlebenswichtigen Fähigkeiten beiträgt (Huizinga, 2009 [1939]). Kleinkinder erfassen ihre Umwelt auf eine spielerische Art und Weise, was sich jedoch im Jugendalter verliert. Stattdessen wird der Lernprozess als eine ernsthafte Angelegenheit deklariert. Erst durch das verstärkte Aufkommen von Videospielen wurden Lernen und Unterhaltung wieder miteinander in Verbindung gebracht (Sheldon, 2012). Lernspiele verbinden dabei die mit dem Spiel einhergehenden Emotionen Freude sowie Spannung bzw. Neugierde, welche als entscheidender Faktor für die Lernmotivation genutzt werden können (Huizinga, 2009 [1939]).

Das game-based Learning basiert nicht darauf, dass eine virtuelle Spieleumgebung geschaffen wird, sondern dass Spielemechanismen beispielsweise im

universitären Kontext integriert werden. Im Folgenden werden typische Mechanismen aus dem Gamedesign präsentiert (Zichermann & Cunningham, 2011):

- Punktesystem: die Vergabe von Punkten (häufig auch Experience Points oder XP genannt) bei erfolgreich gelösten Aufgaben führt zu einer positiven, extrinsisch-motivierten Verstärkung des gewünschten Verhaltens. Darüber hinaus bieten sie dem Nutzer eine direkte Rückmeldung über die erbrachte Leistung.
- Level: eng verbunden mit der Punktevergabe ist das Leveling. Beim Erreichen fest definierter Punktegrenzen steigt der Nutzer ein Level auf. Dadurch wird nicht nur der Fortschritt eines Spielers angezeigt, sondern die geht weitergehendeinher mit einem erhöhten Schwierigkeitsgrad in den folgenden Leveln.
- Achievements: Achievements stellen Erfolge dar, die ein Spieler erreicht hat, welche er auch gegenüber anderen Spielern präsentieren kann. Damit wird einerseits die Sammelleidenschaft angesprochen und andererseits auch Feedback vermittelt.
- Quests: um Punkte oder Achievements zu erhalten, müssen Aufgaben (Quests) gelöst werden. Diese können für sich allein stehen, aber auch in einem größeren thematischen Zusammenhang als Questlinien mit mehreren Quests realisiert werden. Der Spieler muss allein oder in einer Gruppe ein fest definiertes Ziel erreichen. Der Schwierigkeitsgrad der Aufgabe korreliert mit dem Level.
- Ranglisten: ein Ranking anhand der erworbenen Erfahrungspunkte und des Levels bietet dem Spieler einen direkten Vergleich zu seinen Mitspielern. Dies spricht vor allem den Wettbewerbsgedanken vieler Spieler an, die somit durch den Wettkampf um Platz eins der Rangliste motiviert werden.

Aufbruch nach Zyren – Einsatz von Spielemechanismen in der Lehre

In Zusammenarbeit mit Masterstudierenden der Informationswissenschaft in Düsseldorf wurden diese Spielemechanismen auf eine Lehrveranstaltung

übertragen. Basierend auf einer Übung in Form einer Präsenzveranstaltung sowie einer virtuellen Plattform wurde für die Vorlesung Wissensrepräsentation für die Bachelorstudierende des zweiten Semesters dementsprechend ein neues Lehrkonzept entwickelt. Anhand einer spannenden Geschichte werden die Studierenden auf der Suche nach dem Buch des Wissens durch das Land Zyren geleitet. Die Geschichte ist in vier Akte unterteilt, wobei jeder Akt wiederum aus vier Questlinien besteht.

Das Herzstück des Konzepts stellt die Plattform dar, auf welcher die Studierenden Aufgaben lösen können. Gemäß ihrem eigenen Tempo können die Studierenden lernen und ihren aktuellen Wissensstand unabhängig von Zeit und Ort überprüfen. Darüber hinaus werden zusätzliche Aufgaben, sogenannte Sidequests, angeboten, mit denen die Studierenden ihr Verständnis der verschiedenen Vorlesungsthemen freiwillig vertiefen und verbessern können. Dabei wurde auf ein breites Spektrum an Aufgabentypen gesetzt, wie beispielsweise Single-/Multiple-Choice-Tests, Lückentexte und Drag-and-Drop-Aufgaben. Des Weiteren steht am Ende jeder Questlinie eine kurze schriftliche Aufgabe zum Einreichen an, sodass die Studierenden das gelernte Wissen eigenständig noch einmal schriftlich zusammenfassen müssen. Die hohe Anzahl an verschiedenen Questtypen bringt ständige Abwechslung mit sich, was zu einer erhöhten Konzentration bei der Bewältigung führt und Monotonie sowie aufkommende Langeweile vermeiden soll (Domagk & Niegemann, 2009). Alle Aufgaben sind so gestaltet, dass die Studierenden nicht in der Story fortfahren können, solange sie die Quests nicht erfolgreich abgeschlossen haben. Dazu gehört allerdings auch, dass jede Quest beliebig oft wiederholt werden kann, falls man daran scheitert. So besteht die Möglichkeit, verschiedene Strategien und Vorgehensweisen auszuprobieren. Da die meisten Aufgaben automatisch kontrolliert werden – abgesehen von der Abgabeform – erhält der Spieler unmittelbares Feedback, ob sein Lösungsansatz korrekt war, oder ob dieser noch optimiert werden muss.

Einhergehend mit dem Absolvieren von Quests erhält der Studierende Erfahrungspunkte, die nicht nur seinen Fortschritt in der Story anzeigen, sondern auch als Maß für seinen Kenntnisstand zu verstehen sind. Dabei zählt nicht, wie oft er eine Aufgabe wiederholen musste, sondern allein dass er sie bewältigt hat. Der Punkterwerb unterstützt einerseits den Studierenden bei der Fehlererkennung und beim Aufzeigen seiner Stärken sowie Schwächen (Knautz, 2012). Andererseits ist es auch eine Rückmeldung für die Dozierenden und die Übungsgruppenleiter, welche anhand des Levels und der Punkte entnehmen können, wo eventuell thematische Schwierigkeiten bei

den Studierenden auftreten. Beim Erleben der Geschichte steigt das Level des Studierenden und somit der Schwierigkeitsgrad der Aufgaben. Nach Csikszentmihalyi (1985) wird somit ein ausgeglichenes Verhältnis zwischen den Anforderungen und den Fähigkeiten geschaffen, was die Basis für das Flow-Erleben darstellt.

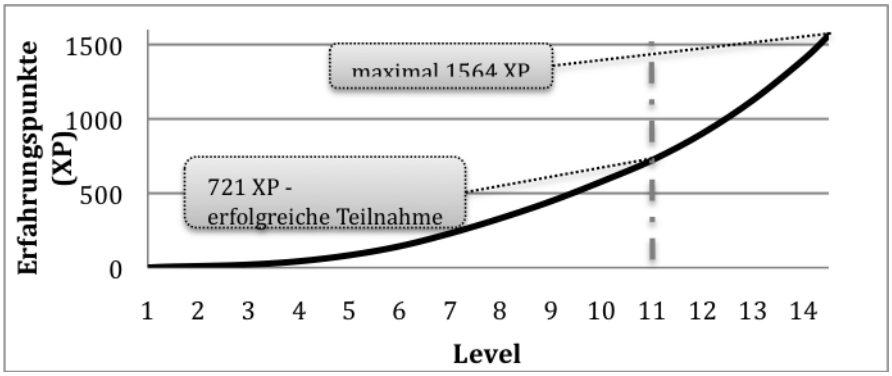

Abbildung 1: Verteilung der Erfahrungspunkte auf das Levelsystem

Anhand des Punktesystems und einer festgelegten Punktezahl lässt sich ferner klar festhalten, wann ein Studierender erfolgreich an der Veranstaltung teilgenommen hat. Die Geschichte geht jedoch über diesen Punkt hinaus, sodass weitere vertiefende Aufgaben angeboten werden. Als Anreiz wurde zudem ein Bonus realisiert, der auf die mündliche Abschlussprüfung angerechnet werden kann, welcher die Studierenden motivieren soll, darüber hinaus das System zu nutzen (vgl. Abb. 1).

Um eine höhere Nutzerbindung an das System zu realisieren, haben die Studierenden die Möglichkeit ihren In-Game-Charakter in verschiedenen Aspekten anzupassen. So können sie das Geschlecht und die Rasse wählen. Sie haben dabei die Wahl zwischen einem Ork, einem Goblin, einem Elfen oder einem Menschen (vgl. Abb. 2). Das fördert einerseits die Identifizierung mit dem eigenem Charakter und andererseits das Zugehörigkeitsgefühl an die Community der Plattform (Martin, 2005; Misoch, 2010). Die Nutzer spielen aber nicht nur voneinander separat, sondern bilden zudem auch feste Lerngruppen, sogenannte Gilden, in denen sich die Spieler innerhalb der Präsenzübungen gegenseitig unterstützen. Dieses Konzept wird auch über die

Plattform unterstützt, da hier ein fester Gildenbereich implementiert wird, in welchem die Spieler miteinander kommunizieren können.

Abbildung 2: Weiblicher Charakter der Rasse Mensch

Die Übung soll vor allem dazu dienen, mit den Studierenden vor Ort Fragen zu klären und die Vorlesungsthemen noch einmal zu vertiefen. Hierbei können auch spontan Probleme angesprochen werden, was auf der Plattform nicht möglich ist. Dabei soll die Präsenzveranstaltung jedoch nicht losgelöst von der Story aufgefasst werden. Stattdessen werden auch in diesem Kontext Spielemechanismen eingesetzt. Einen wichtigen Bestandteil bilden hierbei die Gilden. In diesen festen Teams werden Aufgaben bewältigt, die auch im Wettkampf mit anderen Gilden gelöst werden müssen. Dadurch soll das Zusammengehörigkeitsgefühl gefördert werden, sodass die Studierenden sich auch in anderen Situationen gegenseitig unterstützen und einander helfen. Ein Beispiel für eine Gildenquest, die auch zur Vorbereitung auf die mündliche Prüfung dient, heißt „Das letzte Training“. Die Gilden müssen sich jeweils fünf Fragen zu den aktuellen Vorlesungsthemen und fünf Fragen über den gesamten Lernstoff ausdenken. Nach einer entsprechenden Vorbereitungsphase treten zwei Gilden gegeneinander an und versuchen durch besonders schwierige Fragen ihre Gegner „auszuschalten“. Das Team, welches als Gruppe die meisten Fragen korrekt beantworten konnte, gewinnt den Kampf. Dabei trainiert diese Aufgabe nicht nur die Studierenden knappe und präzise Antworten zu geben, sondern zeigt außerdem auf, wo Wissenslücken vor-

handen sind. Durch das Lösen von Gildenquests sammeln die Teilnehmer Erfahrungspunkte für ihre Gilde, sodass diese im Gilden-Ranking aufsteigen kann. Als Pendant zu den Gildenquests gibt es zusätzlich die Pick-up-Groups. Hierbei werden die festen Lerngruppen aufgebrochen und die Teilnehmer müssen sich in spontan-gebildeten Teams zusammenfinden, um gemeinschaftlich ein Problem zu bewältigen. Auf diese Weise müssen sich die Studierenden nicht nur auf ihr festes Team einstellen, sondern darüber hinaus auch mit anderen Kommilitonen interagieren. Sowohl die Gildenquests als auch die Aufgaben der Pick-up-Groups sind in die Geschichte eingebettet und bilden somit einen direkten Bezug zur Plattform.

Evaluierung des Lehrkonzepts

Die Lehrveranstaltung wurde den Studierenden angekündigt und das modifizierte Lehrkonzept wird im Sommersemester 2013 zum ersten Mal durchgeführt.

Im Vorfeld wurde bereits eine Befragung durchgeführt, in der die generelle Einstellung der Studierenden gegenüber Spielen jeglicher Art erfasst wurde. Die Auswertung lässt deutliche Tendenzen erkennen: knapp 90% aller Befragten spielen in ihrer Freizeit, wovon wiederum etwa 75% unter anderem ihre Zeit mit Videospielen verbringen. Des Weiteren wurde hinsichtlich des neuen Konzepts das Meinungsbild der Studierenden zu dem Einsatz von Gamification in der Hochschullehre ermittelt. Ein Großteil der Studierenden zeigt Interesse an der überarbeiteten Veranstaltung und steht dem positiv gegenüber, das bestehende didaktische Modell zu optimieren. Dabei soll es als eine Alternative zu Dozenten-basierten Lehrveranstaltungen angesehen werden. Hierbei wird vor allem ein Fokus darauf gesetzt, dass anhand von Spielemechanismen bisherige didaktische Konzepte aufgelockert und eine höhere Interaktion der Studierenden gefördert werden.

Die semesterbegleitende Evaluierung setzt verstärkt auf eine Befragung der Beteiligten: Studierende, Dozierende, Übungsgruppenleiter und begleitende Tutoren aus dem Masterstudiengang. Dazu werden verschiedene Fragebögen entsprechend der Personengruppen eingesetzt. In regelmäßigen Abständen werden sowohl quantitative Befragungen als auch qualitative Interviews ge-

führt, um somit ein Meinungsbild über den Verlauf der gesamten Veranstaltung zu erhalten (vgl. Abb. 3).

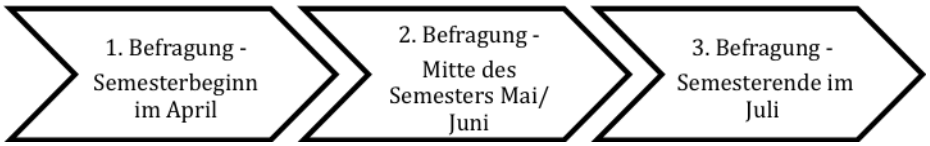

Abbildung 3: Zeitlicher Ablauf der verschiedenen Befragungsphasen

Anhand dieses verteilten Prozesses kann man überprüfen, ob eventuell die Studierenden den Einstieg in die Veranstaltung als besonders schwierig empfinden oder ob die Motivation und das Interesse zeitnah zur Prüfung stark zunehmen. Sollten darüber hinaus massive Probleme mit der Präsenzveranstaltung oder der Plattform auftreten, können diese erkannt und dementsprechend korrigierende Maßnahmen vorgenommen werden. Denn Zweck der Datenerhebung ist es, nicht nur die Akzeptanz des didaktischen Konzepts zu überprüfen, sondern auch die Usability der Plattform bei massiver Nutzung. Dafür werden unter anderem Heuristiken der Software-Evaluation eingesetzt (vgl. Malone, 1980; Epstein, Gelfand, & Lock, 1998).

Des Weiteren sollen unter anderem die folgenden Fragen durch die Evaluierung geklärt werden:

- Wie viele Studierende erledigen nur das Minimum an Anforderungen und wie viele sind motiviert, einen Bonus zu erarbeiten?
- Lässt sich eine Korrelation zwischen guten Prüfungsleistungen und den aktiven Teilnehmern des modifizierten Konzepts feststellen?
- Wie viele Versuche benötigen die Studierenden pro Aufgabe?
- Ändert sich die durchschnittliche Anzahl der Versuche zwischen Main- und Sidequests?
- Wie verhält sich die Anzahl der Studierenden, die direkt beim ersten Mal die meisten Aufgaben korrekt lösen, zu denen, die möglicherweise dazu tendieren sich durchzuklicken?
- Spielen die Studierenden regelmäßig oder bearbeiten sie die Quests eher zusammenhängend ohne Unterbrechungen?

Fazit

Durch den hohen technologischen Standard lassen sich inzwischen andere didaktische Konzepte realisieren als dies noch vor einigen Jahren der Fall war. Ein Beispiel hierfür stellt das Konzept des game-based Learnings dar. Dabei stellt es keine Besonderheit dar, dass Spiele für Lernzwecke eingesetzt werden, sondern dass ein Spiel wie „Die Legende von Zyren“ in eine digitale Umgebung verlagert wird, in welcher die Lernenden sich bewegen und Aufgaben lösen. Ein zentraler Aspekt hierbei ist die unmittelbare Rückmeldung zur ausgeführten Handlung. Diese ermöglicht es dem Spieler, darauf entsprechend zu reagieren, um seine Lernstrategie zu optimieren. Dieser Ablauf von Handlung und Feedback bildet einen Zyklus, der den Lernstoff festigt (Garris, Ahlers, & Driskell, 2002). Durch die Verbindung mit der Suche nach dem Buch des Wissens in *der Legende von Zyren* entsteht daraus eine Variante eines Text-Adventures, wodurch typische Begleitemotionen von Spielen wie Spannung und Freude angesprochen werden, welche als Auslöser für eine intrinsische Lernmotivation dienen. Ziel des gesamten Projekts ist es somit nicht nur die aktive Lernerfahrung zu stärken, sondern die Lernfreude der Studierenden am Thema zu wecken.

Dass das Konzept einer Lehrveranstaltung mit game-based Learning Elementen durchaus diese Ziele erreichen kann, zeigen Beispiele wie Quest to Learn und die Khan Academy. Bei Quest to Learn handelt es sich um eine öffentliche Schule in New York, die ihr gesamtes Curriculum auf game-based Learning aufbaut. Die Khan Academy ist eine Plattform, auf welcher interessierte Nutzer weltweit interaktive und multimediale Lerninhalte nutzen können. Dabei wird auf eine Kombination aus Lehrvideos und Aufgaben mit individuell angepasstem Schwierigkeitsgrad gesetzt. Aber auch im universitären Kontext gibt es Vorbilder für Gamification in der Lehre. Das bekannteste Beispiel hierbei sind die Kurse von Lee Sheldon, der seine Seminare als Multiplayer-Spiele organisiert. Mit ihrem Avatar müssen die Studierenden Quests lösen und Erfahrungspunkte sammeln, um somit den Kurs abzuschließen. Das Feedback der Studierenden sowie die Abschlussergebnisse zeigen deutlich, dass dieses Konzept einen positiven Einfluss auf das Lernverhalten hat (Sheldon, 2010). Ähnliche Ergebnisse bezüglich des Spaßfaktors und der Lernfreude wurden von Ebner und Holzinger (2007) verzeichnet, die ein Online-Spiel für einen Masterkurs des Bauingenieurwesens entwickelten.

Es lässt sich somit abschließend festhalten, dass das spielerische Lernen auch bei jungen Erwachsenen einen positiven Einfluss auf das Lernverhalten hat. Vor allem der Aspekt der Freude wirkt dabei verstärkend. Für das Konzept der *Legende von Zyren* werden voraussichtlich zum Ende des Sommersemesters 2013 Evaluationsergebnisse vorliegen.

Literaturverzeichnis

Artelt, C., Baumert, J., Julius-McElvany, N., & Peschar, J. (2003). Learners for Life – Student Approaches to Learning. Paris. OECD.

Csikszentmihalyi, M. (1985). Das Flow-Erlebnis: Jenseits von Angst und Langeweile: im Tun aufgehen. Stuttgart: Klett-Cotta.

Domagk, S., & Niegemann, H. M. (2009). Emotion, Narration und Lernen mit interaktiven Lernangeboten. In H. W. Giessen (Hrsg.), Emotionale Intelligenz in der Schule. Unterrichten mit Geschichten (S. 39-62). Weinheim / Basel: Beltz.

Ebner, M., & Holzinger, A. (2007). Successful implementation of user-centered game based learning in higher education: An example from civil engineering. Computers & Education, 49(3), 873-890.

Epstein, S. L., Gelfand, J., & Lock, T. E. (1998). Learning Game-Specific Spatially-Oriented Heuristics. Constraints, 3(2/3), 239-253.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, Motivation, and Learning: A Research and Practice Model. Simulation & Gaming, 33 (4), 441-467.

Gaston, J. (2006). Reaching and Teaching the Digital Natives. Library Hi Tech News, 23(3), 12-13.

Gust von Loh, S., & Stock, W. G. (2013). Informationskompetenz als Schulfach? In S. Gust von Loh & W.G. Stock (Hrsg.), Informationskompetenz in der Schule. Ein informationswissenschaftlicher Ansatz (S. 1-20). Berlin / Boston, MA: De Gruyter Saur.

Heinze, N., & Schnurr, J. M. (2010). Informationskompetenz als Baustein für lebenslanges Lernen. In M. H. Breitner, C. Voigtländer, & K. Sohns (Hrsg.), Perspektiven des Lebenslangen Lernens - dynamische Bildungsnetzwerke, Geschäftsmodelle, Trends (pp. 183-193). Berlin: Gito Verlag.

Huizinga, J. (2009 [1939]). Homo ludens. Vom Ursprung der Kultur im Spiel Reinbek: Rowohlt (Original v. 1939).

Knautz, K. (2012). Gamification im Kontext der Vermittlung von Informationskompetenz. In S. Gust von Loh & W.G. Stock (Hrsg.), Informationskompetenz in der Schule. Ein informationswissenschaftlicher Ansatz (S. 223-257). Berlin / Boston, MA: De Gruyter Saur.

Malone, T. W. (1980). What Makes Things Fun To Learn - Heuristics For Designing Instructional Computer Games. In SIGSMALL '80 Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems, S. 162-169.

Martin, J. (2005). Virtually Visual: The Effects of Visual Technologies on Online Identification. In Digital Games Research Conference 2005, Changing Views: Worlds in Play. Retrieved April 08, 2013 from http://www.digra.org/dl/db/06278.08106.pdf

Misoch, S. (2010). Avatare: Spiel(er)figuren in virtuellen Welten. In K.-U. Hugger (Hrsg.), Digitale Jugendkulturen (S. 169-185). Wiesbaden: VS Verlag für Sozialwissenschaften.

Myers, D.G (200). Psychologie. Heidelberg: Springer.

Prensky, M. (2001a). Digital Natives, Digital Immigrants. On The Horizon, 9(5).

Prensky, M. (2001b). Do They Really Think Differently? On The Horizon, 9(6).

Schlag, B. (2009). Lern- und Leistungsmotivation. Wiesbaden: Springer.

Sheldon, L. (2012). The Multiplayer Classroom. Boston: Course Technology.

Zichermann, G. & Cunningham (2011). Gamification by Design. Beijing: O'Reilly.

Was bedeutet 11.882 Bearbeitungen?

Die qualitative Inhaltsanalyse zur Untersuchung des Entstehungsprozesses von Wikipedia-Artikeln

Clemens Roth, Joachim Griesbaum, Ralph Kölle

Stiftung Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
c.roth@fastmail.fm, {griesbau | koelle}@uni-hildesheim.de

Zusammenfassung

Der Wikipedia-Artikel „Deutschland“ ist trotz 11.882 Bearbeitungen kein als exzellent oder lesenswert ausgezeichneter Artikel. Man kann dementsprechend davon ausgehen, dass er im Sinne der Wikipedia-Qualitätskriterien nicht besonders hochwertig ist. Dieses Beispiel dient als Aufhänger für die Frage, ob und inwiefern Zusammenhänge zwischen der Anzahl und Art von Bearbeitungen und der Qualität von Beiträgen existieren. Was bedeutet eine Bearbeitung eines Wikipedia-Artikels hinsichtlich der Qualität des Beitrags und welche Aspekte sind bei der Untersuchung dieses Zusammenhangs wichtig? Um diese Frage zu adressieren, wurde eine explorative Studie in Form einer qualitativen Inhaltsanalyse von sechs Wikipedia-Artikeln durchgeführt. Der nachfolgende Artikel stellt den methodischen Ansatz dieser Studie sowie deren Ergebnisse dar.

Abstract

What does high edit count mean and which aspects are important when analyzing the evolution of Wikipedia articles? This paper discusses methods and approaches to investigate relations between the evolution of Wikipedia articles and their quality. Results of a first exploratory analysis which focuses on a qualitative in-depth categorization of an article's history are presented.

Einleitung

Die Wikipedia ist wohl die größte (Online-) Enzyklopädie (am 29.03.2013 weist die deutschsprachige Version über 1,5 Mio Einträge auf) und gehört zu den meistbesuchten Seiten im Internet (vgl. Alexa 2013). Die Artikel werden von freiwilligen, teils anonymen Benutzern geschrieben, können von jedem bearbeitet werden und sind frei verfügbar. (vgl. Wikipedia 2013a)

Die große Popularität und die freie Verfügbarkeit aller Inhalte stellen Gründe dafür dar, dass die Wikipedia in vielfältiger Hinsicht erforscht wird. Einen Schwerpunkt bildet dabei die Frage der Qualität der Inhalte. Neben Untersuchungen, die einen Schwerpunkt auf artefaktzustandsbezogene Attribute wie etwa Beitragslänge, Schreibstil, Objektivität oder Anzahl vorhandener Links legen, z.B. Lewandowski & Spree (2011), Yaari et al. (2011), existieren auch Ansätze, welche dabei den Entstehungsprozess der Beiträge fokussieren, z.B. Liu & Ram (2011), Kane (2011), Stein & Hess (2008). In diesen Bereich ordnet sich die hier dargestellte Untersuchung mit ein. Die Kernfrage ist, zu eruieren, welche Arten von Bearbeitungen an Wikipedia-Artikeln durchgeführt werden, wie häufig diese vorkommen und ob sich diesbezüglich Unterschiede zwischen Artikeln verschiedener Qualitätsstufen identifizieren lassen.

Der Wikipedia-Artikel „Deutschland“ eignet sich, um das Erkenntnisinteresse zu veranschaulichen. Dieser weist zwar mit 11.882 Bearbeitungen die höchste Anzahl an Bearbeitungen in der deutschen Wikipedia auf (Wikipedia 2013b: Stand: 27.1.2013), gleichzeitig ist er nicht mehr als lesenswert oder exzellent ausgezeichnet. Er wird also aktuell, zumindest wenn man die Auszeichnungsverfahren der Wikipedia selbst (Wikipedia 2013c) als Vergleichsmaßstab nimmt, trotz der besonders vielen Bearbeitungen nicht als besonders hochwertiger Artikel betrachtet.

Anhand einer qualitativen Inhaltsanalyse von sechs Wikipedia-Artikeln verschiedener Qualitätsstufen soll in diesem Beitrag ein konkreter Ansatz zur entstehungsprozessbezogenen Analyse vorgestellt und diskutiert werden.

Hierzu wird zunächst das Konzept der Versionshistorie in der Wikipedia skizziert. Anschließend werden methodische Grundlagen der Inhaltsanalyse angeführt. Auf dieser Basis wird ein kurzer State of the Art zu quantitativen und qualitativen Untersuchungen im Themenfeld gegeben. Schließlich werden das eigene Vorgehen und zentrale Ergebnisse der Untersuchung geschil-

dert. Abschließend wird die Untersuchung eingeordnet und ein Ausblick auf mögliche weitere Untersuchungen gegeben.

Das Konzept der Versionshistorie bei der Wikipedia

Eines der wichtigsten Charakteristika der Wikipedia ist die lückenlose Speicherung aller Änderungen an Artikeln in der Versionshistorie. Sie ermöglicht ein genaues Nachvollziehen aller Bearbeitungen und so des gesamten Entstehungsprozesses.

Nach jeder Bearbeitung eines Artikels, d.h. immer dann, wenn ein Benutzer auf Speichern klickt, wird eine neue Version inklusive verschiedener Metadaten, wie Autor, Datum und Kommentar, in der Versionshistorie angelegt (vgl. Wikipedia 2013d). Welche Änderung er tatsächlich vorgenommen hat, ist nicht ersichtlich. Die Quantität der Versionen erlaubt damit zunächst keine Rückschlüsse hinsichtlich der qualitativen Entwicklung eines Artikels. Auch bei Änderungen, die den Artikel nicht inhaltlich betreffen, werden neue Versionen erzeugt. Beispiele hierfür sind Vandalismus und Revertierungen, also das vorsätzliche Verfälschen und Beschädigen von Artikeln und das anschließende Wiederherstellen von vorhergehenden Versionen sowie Wikipedia-spezifische Aktivitäten, wie das Hinzufügen von Kategorien oder Links zu anderen Sprachversionen. Des Weiteren werden von einem Autor oft mehrere Versionen hintereinander erstellt. Diese können zu einem Bearbeitungsschritt zusammengefasst werden.

Zusammenfassend lässt sich konstatieren, dass sich aus der Versionsanzahl eines Artikels zunächst nur die Quantität der Editierungen ablesen lässt. Es können keine weiteren Aussagen über die Entstehung eines Artikels und seine Qualitätsentwicklung getroffen werden. Für derartige Einschätzungen ist die Durchführung von inhaltsanalytischen Untersuchungen unabdingbar.

Inhaltsanalyse

Nach Mayring ist die Inhaltsanalyse ein sehr flexibles Werkzeug um „[...] Material, das aus irgendeiner Art Kommunikation stammt“ (Mayring 2010:11), zu analysieren. Wobei sie „[...] nicht nur Inhalte der Kommunikation zum Gegenstand hat [...]“ (Mayring 2010:13), sondern auch als eine kategoriengeleitete Textanalyse gesehen werden kann.

Die Grundlage der Inhaltsanalyse ist ein Kategoriensystem, das auf ein bestimmtes Material, zum Beispiel Foren-Diskussionen oder auch Zeitungsartikel, angewendet wird. Dabei wird zwischen deduktiver und induktiver Kategorienbildung unterschieden. Deduktiv bedeutet, dass Kategorien a priori aus der Theorie oder aus dem Forschungsinteresse abgeleitet und begründet werden. Bei der induktiven Vorgehensweise werden die Kategorien direkt aus dem zu untersuchenden Material extrahiert. (vgl. Mayring 2010:49f.) Reinhoffer (2008:127) weist daraufhin, dass deduktive Ansätze zwar eine größtmögliche Systematik und ein regelgeleitetes Vorgehen bieten, aber manchmal nicht komplett offen und nachvollziehbar sind. Zudem ist es möglich, dass bestimmte Aspekte des Materials übersehen werden, da sie vorher nicht bekannt waren. Induktive Vorgehensweisen bieten hingegen eine sehr große Offenheit und die Möglichkeit, alle Varianzen des Materials im Kategoriensystem abzubilden.

Bei der Entwicklung eines Kategoriensystems muss zudem entschieden werden, ob die Texte quantitativ oder qualitativ kategorisiert werden sollen. Des Weiteren müssen klare Definitionen und Regeln für die einzelnen Kategorien festgelegt und möglichst Beispiele zur Illustration gefunden werden.

Inhaltsanalysen von Wikipedia-Artikeln

Um den tatsächlichen Entwicklungsprozess von Wikipedia-Artikeln im Sinne der Zusammenarbeit verschiedener Benutzer analysieren zu können, ist es sinnvoll, zunächst die Versionshistorie zu filtern. Dabei werden nicht relevante Versionen entfernt und aufeinanderfolgende Versionen eines Autors zu Bearbeitungsschritten zusammengefasst. Die Anzahl der so entfernten Versionen schwankt zwischen verschiedenen Artikeln sehr stark, da insbesondere

die Anzahl der als Vandalismus einzustufenden Versionen unterschiedlich ist und auch die Anzahl der durch Zwischenspeicherungen entstandenen Versionen stark von dem individuellen Verhalten der Autoren abhängt. In der vorliegenden Untersuchung liegt die Anzahl der entfernten Versionen zwischen 219 und 599 Versionen, das entspricht einem prozentualen Anteil zwischen 50,8% und 69% an allen Editierungen.

Anschließend können die Bearbeitungsschritte anhand des vorher definierten Kategoriensystems kategorisiert werden. Dazu bietet sich die Verwendung einer Diff-Funktion an. Sie visualisiert die Änderungen der Bearbeitungsschritte eines Artikels, indem die Unterschiede zweier aufeinanderfolgender Versionen textuell gekennzeichnet werden.

Auf diese Weise lassen sich zum Beispiel Aussagen über die Häufigkeit und Verteilung verschiedener Aktionen treffen (Roth 2012), kulturelle Unterschiede im Editierverhalten untersuchen (Pfeil et al. 2006) oder Rollen und Kollaborationsmuster definieren, die auf den Aktivitäten der Autoren basieren (Liu & Ram 2011). Die Schwierigkeit bei der Entwicklung von Kategorien und auch bei der späteren Codierung liegt darin, dass meist nicht bekannt ist, welche Intention einer Änderung zugrunde lag. Bei der induktiven Entwicklung des Kategoriensystems werden bestimmte Aktionen aus den Unterschieden zwischen zwei Text-Versionen abgeleitet und in Kategorien abstrahiert. Bei dem Kategorisieren bzw. Codieren handelt es sich schließlich um die Interpretation von Unterschieden zwischen zwei Text-Versionen auf Basis festgelegter Regeln. Die Unterschiede werden anhand des Kategoriensystems codiert und repräsentieren dann Aktionen von Autoren.

Qualitative oder quantitative Datenerhebung?

Insgesamt werden im Rahmen von Inhaltsanalysen bestimmte Sachverhalte quantifiziert. Bei der Konzeption einer Inhaltsanalyse ist abzuwägen, ob die Codierung von Menschen durchgeführt werden muss oder automatisiert erfolgen kann. Dies ist davon abhängig, wie bestimmte Sachverhalte abstrahiert werden und welche Regeln und Indikatoren zur Codierung herangezogen werden.

Die folgende Darstellung von Untersuchungen im Themenfeld verdeutlicht die Unterschiede beider Vorgehensweisen.

Übersicht zu inhaltsanalytischen Untersuchungen

Die nachfolgende Darstellung erhebt keinen Anspruch auf Vollständigkeit. Sie dient primär dazu, das Vorgehen der eigenen Untersuchung einzuordnen.

In qualitativen Studien zur Versionshistorie von Wikipedia-Artikeln werden die verschiedenen Aktionen der Benutzer von Menschen interpretiert bzw. codiert. Dadurch kann meist auf die Intention und somit auf die tatsächlich durchgeführte Aktion geschlossen werden.

Pfeil et al. (2006) zeigen in einer qualitativen Untersuchung mehrerer Sprachversionen eines Artikels, dass sich kulturelle Unterschiede im Editierverhalten von Wikipedia-Autoren beobachten lassen. Dabei setzen sie ein Kategoriensystem ein, das auch zwischen inhaltlichen und formalen Änderungen unterscheidet. Kane (2011) führt zunächst eine qualitative Untersuchung der Versionshistorie eines Artikels durch und identifiziert viele verschiedene Arten von Editierungen. Insbesondere das sogenannte Content Shaping, sinngemäß das Formatieren, Umstrukturieren und Überarbeiten von fremdem Inhalt ist von großer Bedeutung. Weitere qualitative Untersuchungen der Versionshistorien liefern Pentzold (2007), Kallass (2008) und Swarts (2009). Pentzold (2007) weist in einer Diskursanalyse von zwei Artikeln diskursive Regime im Sinne Michel Foucaults nach, während Kallass (2008) Aspekte des kollaborativen Schreibens anhand der Entwicklung von zwei Wikipedia-Artikeln untersucht. Swarts (2009) erforscht wiederum die kollaborative Konstruktion von Fakten und kommt zu dem Schluss, dass sie wie alle wissenschaftlichen Fakten in einem Prozess entstehen.

Diese qualitativen Kategoriensysteme haben den Vorteil, dass detailliert unterschieden werden kann, ob ein Bearbeitungsschritt den Inhalt verändert hat oder nicht. So kann der Entwicklungsprozess von Wikipedia-Artikeln sehr genau analysiert und nachvollzogen werden. Die Bearbeitungsschritte müssen von Menschen und unter Einbezug verschiedener Kontext-Informationen intellektuell interpretiert und codiert werden. Ein zentraler methodischer Aspekt ist dabei die Frage der Sicherstellung der Intersubjektivität der intellektuellen Codierung. Weitergehend stellt die mangelnde Automatisierbarkeit einen zentralen Nachteil dar. Dadurch wird der Aufwand bei der Datenerhebung meist sehr hoch und eine große Stichprobe kann nur mit erheblichem Ressourcenaufwand untersucht werden.

Quantitative Untersuchungen

Bei quantitativen Untersuchungen von Wikipedia-Artikeln und deren Versionshistorie werden die Benutzeraktivitäten auf Basis formaler Regeln abstrahiert. Das heißt, die Änderungen werden mit einer Diff-Software automatisch erkannt und direkt kategorisiert.

Stein & Hess (2008) untersuchen die Entwicklung exzellenter Artikel anhand von Editierungsquantitäten. Sie verfolgen die Frage, ob die Anzahl der Autoren oder eher die Qualität einzelner Autoren bei der Entstehung exzellenter Artikel entscheidend ist. Dafür bewerten sie die Fähigkeit eines Autors anhand der Anzahl von ihm getätigter Editierungen exzellenter Artikel und der Anzahl exzellenter Artikel, die von ihm editiert wurden. Es gilt die Grundannahme: Je höher der jeweilige Anteil an seinen gesamten Editierungen bzw. insgesamt editierten Artikeln, als desto fähiger wird der Autor eingestuft. Es wird dabei nicht berücksichtigt, welche inhaltlichen Aktionen er vollzieht oder wie umfassend diese sind. Anhand des Anteils der fähigen Autoren, die an einem Artikel mitwirken, wird auf die Qualität des Artikels geschlossen. Die Autoren kommen zu dem Schluss, dass nicht die reine Anzahl der Autoren entscheidend für die Entstehung exzellenter und lesenswerter Artikel ist, sondern die Fähigkeit einzelner Autoren.

Ebenfalls einem quantitativen Ansatz folgend, analysieren Liu & Ram (2011) 1.600 Artikel der englischen Wikipedia. Dabei wird jede Änderung eines Benutzers automatisch erfasst und ausgezählt. Es wird zwischen den drei Aktivitäten Einfügen, Verändern und Löschen unterschieden. Diese Aktivitäten werden dann auf die Einheiten Satz, Link und Referenz bezogen. Zusätzlich gibt es die Aktion Revertierung. Es werden also alle Sätze von zwei Versionen formal miteinander verglichen und für jeden Satz abgeleitet, ob er neu hinzugefügt, verändert, gelöscht oder nicht verändert wurde. In verschiedenen weiteren Analysen werden anschließend Kollaborationsmuster bzw. -rollen, wie *All-round Contributor*, *Watchdog* und *Content Justifier*, entwickelt, um die Artikelqualität vorherzusagen. Die Autoren kommen zu dem Schluss, dass die von ihnen entwickelten Kollaborationsmuster eine gute Möglichkeit bieten, Artikelqualität vorherzusagen und dass die Artikelqualität von den Autoren und deren Rolle abhängt.

Der Vorteil quantitativer Inhaltsanalysen ist, dass sehr große und somit repräsentative Stichproben untersucht werden können. Gleichzeitig können durch die große Abstraktion die Aktionen nicht so detailliert erfasst werden wie in qualitativen Erhebungen. Es können keine inhaltlichen Aspekte be-

rücksichtigt werden. Die Aktion kann letztlich neuen Inhalt hinzufügen, eine inhaltliche Ergänzung, eine Korrektur auf inhaltlicher oder formaler Basis oder auch eine sprachlich-stilistische Änderung sein.

Untersuchungsdesign und Durchführung

Da das Ziel der hier dargestellten Untersuchung zunächst darin bestand, einen explorativen Einblick bezüglich der Anzahl und auch der Art der Bearbeitungen zu gewinnen, wurde ein qualitatives Untersuchungskonzept gewählt. Dabei werden im Rahmen der qualitativen Inhaltsanalyse die Versionshistorien von sechs Artikeln verschiedener Qualität zunächst wie oben beschrieben gefiltert, so dass nur noch inhaltlich relevante Bearbeitungsschritte enthalten sind. Anschließend werden diese anhand eines Kategoriensystems von mehreren Menschen codiert. Das Ziel der Analyse ist es, die kollaborative Entwicklung der Inhalte der Wikipedia-Artikel möglichst umfassend und detailliert zu beschreiben und zu prüfen, ob sich Unterschiede in der Entstehung von Artikeln verschiedener Qualitätsstufen identifizieren lassen. Ein besonderer Fokus liegt dabei auf der Unterscheidung zwischen inhaltlichen und nicht-inhaltlichen Änderungen.

Das Kategoriensystem wurde zunächst deduktiv aus der Theorie der Wissenskonstruktion in Wikis nach Cress & Kimmerle (2008), den Ausführungen von Schmalz (2007) zu Organisationsprinzipien in Wikis und der Untersuchung zu kulturellen Unterschieden im Editierverhalten von Wikipedia-Autoren von Pfeil et al. (2006) abgeleitet. Anschließend wurde es induktiv am Material getestet und weiterentwickelt. Es besteht aus drei Oberkategorien:

A. Inhaltliche Änderungen
B. Formale, nicht-inhaltliche Änderungen
C. Sonstiges

Die Oberkategorie *Inhaltliche Änderungen* fasst alle Aktionen, die inhaltliche Aussagen bzw. den semantischen Gehalt eines Artikels verändern. Dazu gehört das Hinzufügen von neuem Inhalt, das Ergänzen und Spezifizieren

sowie das Korrigieren und Aktualisieren von bestehendem Inhalt sowie das gezielte, begründete Löschen von Inhalt.

Die zweite Oberkategorie, *Formale, nicht-inhaltliche Änderungen*, beinhaltet alle Aktionen, die Rechtschreibung, Sprache und Stil, Struktur sowie die Formatierung und das Layout betreffen. Entscheidend ist, dass inhaltliche Aspekte des Artikels nicht verändert werden.

Die dritte Oberkategorie *Sonstiges* enthält alle weiteren Aktionen, die den Artikel und seine Form nicht beeinflussen. Dazu gehört beispielsweise Vandalismus, das Wiederherstellen von Versionen oder das Hinzufügen von Kategorien oder Links zu anderen Sprachversionen.

Dieses Kategoriensystem ermöglicht es, die Häufigkeit und Verteilung der einzelnen Aktionen, die bei der Bearbeitung von Wikipedia-Artikeln auftreten, detailliert zu analysieren.

Für die Kategorisierung der einzelnen Bearbeitungsschritte wurde ein web-basiertes Tool entwickelt, dass dem Codierer die Bearbeitungsschritte nacheinander anzeigt und es gestattet, die Kategorien zu jedem Schritt direkt über ein Formular einzutragen. Alle Eingaben werden in einer Datenbank gespeichert. Die nachfolgende Abbildung zeigt das Werkzeug in der Übersicht.

Abbildung 1: Das Web-Interface für die Kategorisierung.

Von den über 1,5 Million Artikeln der deutschsprachigen Wikipedia sind 2195 als exzellent (vgl. Wikipedia 2013e) und 3539 als lesenswert (vgl. Wikipedia 2013f) ausgezeichnet (Stand: 15.4.2013). Die Verwendung dieser Qualitätsstufen hat den Vorteil, dass bei ausgezeichneten Artikeln von einem tatsächlichen qualitativen und formalen Mindeststandard ausgegangen werden kann. Für die vorliegende Untersuchung wurde eine Stichprobe von sechs Wikipedia-Artikeln verschiedener Qualität gewählt (ohne Auszeichnung: Kobe Bryant, Curaçao; Lesenswert: Chaosforschung, Lee Harvey Oswald; Exzellent: Tempel von Abu Simbel, Gesundheitssystem der Vereinigten Staaten). Die Stichprobe besteht aus zwei exzellenten, zwei lesenswerten und zwei nicht-ausgezeichneten Artikeln, die zufällig ausgewählt wurden und jeweils mindestens 400 Versionen aufweisen. Sie wurden am 7.7.2012 heruntergeladen.

Die Versionshistorien der Artikel wurden gefiltert und anschließend mit Hilfe des Web-Interfaces codiert. Um die Intersubjektivität zu gewährleisten, wurde jeder Artikel von zwei Codierern, die vorab geschult wurden, kategorisiert. Die Codierungen der verschiedenen Codierer wurden anschließend anhand des Reliabilitätskoeffizienten von Holsti (vgl. Früh 2007:190) verglichen. Die Werte reichen von 0,7 (Curaçao) über 0,74 (Lee Harvey Oswald) und 0,75 (Gesundheitssystem der Vereinigten Staaten) bis hin zu 0,76 (Kobe Bryant, Chaosforschung und Tempel von Abu Simbel). Um die Kategorisierungen zu konsolidieren, wurde anschließend gemeinsam mit den Codierern eine eindeutige Einordnung für die unterschiedlichen Kategorisierungen festgelegt. Somit konnte sichergestellt werden, dass die Kategorien nachvollziehbar, verlässlich und reliabel sind. Nachfolgend werden die Ergebnisse der Untersuchung zusammengefasst. Eine ausführliche Darstellung des Untersuchungsdesigns und der Ergebnisse findet sich in Roth (2012).

Ergebnisse

Die Anzahl der Codierungen der einzelnen Artikel schwankt zwischen 266 und 533 Codierungen. Dabei überwiegen in den beiden nicht-ausgezeichneten Artikeln die Kategorien der Oberkategorie *Inhaltliche Änderungen* mit 48,3% und 48,1% gegenüber 38,5% und 30,1% der Oberkategorie *Formale, nicht-inhaltliche Änderungen.* In den vier ausgezeichneten

Artikeln zeigt sich ein umgekehrtes Bild. Hier überwiegen die Kategorien der Oberkategorie *Formale, nicht-inhaltliche Änderungen* mit durchschnittlich 46,1% gegenüber durchschnittlich 32,8% der Oberkategorie *Inhaltliche Änderungen.*

Tabelle 1: Übersicht der Codierungen ohne die Oberkategorie 'Sonstiges'.

	Kobe Bryant	Curaçao	Chaos-forschung	Lee Harvey Oswald	Tempel von Abu Simbel	Gesund-heitssystem der Vereinigten Staaten
Gesamt	418	266	418	365	293	533
Oberkategorie *Inhaltliche Änderungen*						
Neuer, selbstständiger Inhalt	35 8,4%	26 9,8%	23 5,5%	31 8,5%	28 9,6%	42 7,9%
Ergänzung, Spezifizierung	76 18,2%	53 19,9%	91 21,8%	49 13,4%	50 17,1%	70 13,1%
Inhaltliche Korrektur, Aktualisierung	71 17 %	32 12%	9 2,2%	10 2,7%	10 3,4%	25 4,7%
Löschung	20 4,8%	17 6,4%	24 5,7%	22 6%	12 4,1%	31 5,8%
Oberkategorie Gesamt	202 48,3%	128 48,1%	147 35,2%	112 30,7%	100 34,1%	168 31,5%
Oberkategorie *Formale, nicht-inhaltliche Änderungen*						
Sprachlich-stilistische Überarbeitung	49 11,7%	27 10,2%	52 12,4%	56 15,3%	29 9,9%	89 16,7%
Tippfehler, Rechtschreibung, Interpunktion, Grammatik	60 14,4%	32 12,0%	86 20,6%	76 20,8	24 8,2%	102 19,1
Struktur	8 1,9%	2 0,8%	12 2,9%	11 3,0%	12 4,1%	28 5,3%
Formatierung, Layout	44 10,5%	19 7,1%	35 8,4%	31 8,5%	46 15,7%	71 13,3%
Oberkategorie Gesamt	161 38,5%	80 30,1%	185 44,3%	174 47,7%	111 37,9%	290 54,4%

Die Verteilungen der einzelnen Kategorien der Oberkategorie *Inhaltliche Änderungen* sind in den meisten Fällen relativ ähnlich. Die Kategorien *Neuer, selbstständiger Inhalt, Ergänzung, Spezifizierung* und *Löschung* weisen anteilige Häufigkeiten von durchschnittlich 8,3% (*Neuer, selbstständiger Inhalt),* 17,2% (*Ergänzung, Spezifizierung*) und 5,5% (*Löschung*) auf. Lediglich die Kategorie *Inhaltliche Korrektur, Aktualisierung* weist stark schwankende anteilige Werte von 2,2% bis 17% auf.

In der Oberkategorie *Formale, nicht-inhaltliche Änderungen* sind hingegen größere Unterschiede zu beobachten. Lediglich die Kategorie *Sprachlich-stilistische Überarbeitung* ist ausgeglichen und weist durchschnittlich 12,7% auf. Die weiteren Kategorien schwanken zwischen 8,2% und 20,8% (Kategorie *Tippfehler, Rechtschreibung, Interpunktion, Grammatik*), 0,8% und 5,3% (Kategorie *Struktur*) sowie 7,1% und 15,7% (Kategorie *Formatierung, Layout*).

Aus diesen unterschiedlichen Verteilungen lassen sich die folgenden in Verbindung zur Artikelqualität setzen. Drei der vier ausgezeichneten Artikel weisen in der Kategorie *Tippfehler, Rechtschreibung, Interpunktion, Grammatik* mit 19,1%, 20,6% und 20,8% einen höheren Wert als die nicht-ausgezeichneten Artikel mit 12% und 14,4% auf.

Auch in der Kategorie *Struktur* zeigt sich eine steigende anteilige Häufigkeit von 1,9% und 0,8% in den nicht-ausgezeichneten Artikeln über 2,9% und 3% in den lesenswerten bis zu 4,1% und 5,3% in den exzellenten Artikeln.

Die anteiligen Häufigkeiten der Kategorie *Formatierung, Layout* sind insbesondere in den beiden exzellenten Artikeln mit 15,7% und 13,3% recht hoch. Die Artikel der Qualitätsstufe Lesenswert liegen bei 8,4% und 8,5% und die nicht-ausgezeichneten Artikel bei 7,5% und 10,5%.

Die Ergebnisse sind zwar nicht verallgemeinerbar, weisen aber darauf hin, dass eher ein hoher Anteil an nicht-inhaltlichen Modifikationen ein Indikator für als qualitativ hochwertig gekennzeichnete Artikel darstellen könnte. Dies geht konform mit den Ergebnissen von Kane (2011) und stellt ein eher kontraintuitives Resultat dar. Das Ergebnis ist wohl auch von der Artikelauswahl der Untersuchung abhängig, dennoch ist es für die weitere Forschung nicht uninteressant, da sich gerade Modifikationen hinsichtlich Rechtschreibung, Struktur, Formatierung und Layout automatisch analysieren lassen.

Des Weiteren lässt sich beobachten, dass insbesondere die ausgezeichneten Artikel eine höhere Autorenaktivität aufweisen. Das heißt, dass einzelne Autoren einen Artikel mehrmals bearbeitet haben. In den nicht ausgezeichne-

ten Artikeln nehmen die Autoren hingegen tendenziell nur sehr wenige Änderungen vor. Dies entspricht den Ergebnissen von Stein & Hess (2008) bzw. Liu & Ram (2011), dass eher einzelne Autoren ausschlaggebend für die Qualität sind.

Fazit und Ausblick

Die Ausführungen haben gezeigt, dass es nicht sinnvoll ist, alleine aus der Versionsanzahl von Wikipedia-Artikeln belastbare Aussagen über die qualitative Entwicklung von Wikipedia-Artikeln zu treffen. 11.882 Bearbeitungen bedeuten nur, dass der Artikel Deutschland 11.882 Mal bearbeitet wurde. Welche Änderungen vorgenommen wurden, ist nicht ersichtlich. Um die Entstehung von Wikipedia-Artikeln zu untersuchen, müssen demnach detailliertere Analysen herangezogen werden. Dabei muss genau geprüft werden, welche Teile des Materials interessant sind und mit welcher Methodik stichhaltige Daten erhoben werden können.

Die Inhaltsanalyse ist ein Werkzeug, das sich aufgrund seiner Flexibilität gut eignet, um sowohl qualitative als auch quantitative Analysen der Versionshistorie von Wikipedia-Artikeln durchzuführen. Es ist jedoch von besonderer Bedeutung, dass im Rahmen der Konzeption der Forschungsarbeit genau geplant wird, wie die Änderungen erhoben bzw. codiert werden und wie abstrakt das Kategoriensystem gestaltet wird.

Für die vorliegende Untersuchung lässt sich diesbezüglich folgende Einschätzung vornehmen. Durch die zweifache Codierung durch verschiedene Personen und die anschließende Angleichung ist davon auszugehen, dass die erhobenen Daten im Sinne der Intercoderreliabilität reliabel sind. Aussagen über die Validität der Ergebnisse lassen sich hingegen nur schwer treffen. Legt man jedoch das Außenkriterium an, also einen Vergleich mit den Ergebnissen ähnlicher Studien, kann auch die Validität angenommen werden. Insbesondere Liu & Ram (2011) und Stein & Hess (2008) kommen zu vergleichbaren Ergebnissen.

Wie verschiedene andere Forschungsarbeiten bedient sich die vorliegende Arbeit qualitativer Methoden zur Datenerhebung. Durch den explorativen Charakter der Untersuchung und der tentativen Art der Ergebnisse ist sie teilweise als bestätigend, z.B. in Bezug auf die Unterscheidung zwischen

paralleler und nicht-paralleler Entwicklung von Artikel und Artikelgegenstand von Kallass (2008), aber vor allem als Vorarbeit, sozusagen als Anstoß für verschiedene weitere Forschungsansätze zu sehen.

Es konnten beispielsweise mehrere Unterschiede in der Entwicklung von Artikeln verschiedener Qualitätsstufen identifiziert werden. Ein möglicher weiterer Forschungsansatz wäre demnach zu prüfen, ob sich die Ergebnisse mit repräsentativen Stichproben reproduzieren lassen und somit als Indikator für eine automatische Qualitätserkennung verwendet werden können.

Zusätzlich wäre es denkbar, die erhobenen Daten noch tiefgehender zu analysieren. Ein möglicher Ansatz wäre beispielsweise Frage, ob sich Bearbeitungen verschiedener Kategorien auf einzelne Autoren konzentrieren und ob sich so differenzierte Rollen identifizieren lassen. Ein weiterer Ansatz könnte die zeitliche Entwicklung von Wikipedia-Artikeln sein und die Frage, ob beispielsweise Bearbeitungsphasen auftreten oder bestimmte Bearbeitungen auf andere folgen. Zudem wäre in diesem Kontext interessant, den Zeitpunkt der Auszeichnung als exzellent oder lesenswerter Artikel einzubeziehen und beispielsweise zu untersuchen, wie sich Artikel vorher und nachher entwickeln.

Um zu prüfen, ob es sich die Ergebnisse auf andere Kontexte als Wikipedia übertragen lassen, wäre es interessant, die Untersuchung anhand anderer Wikis durchzuführen.

Aus pragmatischer Sicht könnte die Erkenntnis, dass qualitativ hochwertige Artikel offenbar von einzelnen Autoren häufiger bearbeitet wurden, dazu verwendet werden, die Qualität von Wikipedia-Artikeln zu steigern, indem die Autoren aufgerufen werden, einzelne Artikel häufiger zu bearbeiten. Ob eine solche Initiative funktioniert, bleibt natürlich offen.

Des Weiteren könnte das entwickelte Kategorisierungs-Tool für das Crowdsourcing von Kategorisierungsarbeiten verwendet werden. Auf diese Weise könnten auch qualitative Inhaltsanalysen repräsentative Stichproben erreichen und beispielsweise die Ergebnisse der vorliegenden Untersuchung quantitativ überprüft werden.

References/Literaturverzeichnis

Alexa (2013): Top Sites. The top 500 sites on the web. Retrieved March 31, 2013 from http://www.alexa.com/topsites

Cress, Ulrike; Kimmerle, Joachim (2008): A systemic and cognitive view on collaborative knowledge building with wikis. In: International Journal of Computer-Supported Collaborative Learning 3 (2), S. 105–122.

Früh, Werner (2007): Inhaltsanalyse. 6., überarbeitete Auflage. Konstanz: UVK Verlagsgesellschaft mbH.

Kallass, Kerstin (2008): Artikelentstehung in der Wikipedia. Zu Textkonstitutionsmustern und Schreibprozessen bei Wikipedia-Artikeln. In: Berichte des DFG-Forschungprojekts "Netzwerkkommunikation im Internet", Nr. 08 (3). Retrieved March 31, 2013 from http://www.netzwerke-im-internet.de/fileadmin/downloads/Bericht2008-3-Wikipedia.pdf

Kane, Gerald C. (2011): A multimethod study of information quality in wiki collaboration. In: ACM Transactions on Management Information Systems 2 (1).

Lewandowski, Dirk & Spree, Ulrike (2011): Ranking of Wikipedia articles in search engines revisited: Fair ranking for reasonable quality? In: Journal of the American Society for Information Science, 62(1):117-132.

Liu, Jun; Ram, Sudha (2011): Who does what: Collaboration patterns in the wikipedia and their impact on article quality. ACM Transactions on Management Information Systems 2 (2).

Mayring, Philipp (2010): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 11., aktualisierte und überarbeitete Auflage. Weinheim und Basel: Beltz Verlag.

Pentzold, Christian (2007): Wikipedia. Diskussionsraum und Informationsspeicher im neuen Netz. München: Verlag Reinhard Fischer (Internet research, 28).

Pfeil, Ulrike; Zaphiris, Panayiotis; Ang, Chee Siang (2006): Cultural Differences in Collaborative Authoring of Wikipedia. Journal of Computer-Mediated Communication 12 (1), S. 88–113.

Reinhoffer, Bernd (2008): Lehrkräfte geben Auskunft über ihren Unterricht. Ein systematisierender Vorschlag zur deduktiven und induktiven Kategorienbildung in der Unterrichtsforschung. In: Philipp Mayring und Michaela Gläser-Zikuda (Hg.): Die Praxis der Qualitativen Inhaltsanalyse. 2. Auflage. Weinheim und Basel: Beltz Verlag, S. 123–141.

Roth, Clemens (2012): Kollaborative Wissenskonstruktion am Beispiel von Wikipedia. Magisterarbeit Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim. (Veröffentlichung auf HilDok, dem Publikationsserver

der Universität Hildesheim (http://opus.bsz-bw.de/ubhi/index.php?la=de), wird angestrebt.)

Schmalz, Jan Sebastian (2007): Zwischen Kooperation und Kollaboration, zwischen Hierarchie und Heterarchie. Organisationsprinzipien und Strukturen von Wikis. In: kommunikation@gesellschaft 8. Retrieved March 31, 2013 from http://www.soz.uni-frankfurt.de/K.G/B5_2007_Schmalz.pdf

Stein, Klaus; Hess, Monica (2008): Viele Autoren, gute Autoren? Eine Untersuchung ausgezeichneter Artikel in der deutschen Wikipedia. In: Paul Alpar und Steffen Blaschke (Hg.): Web 2.0 - Eine empirische Bestandsaufnahme. Wiesbaden: Vieweg+Teubner, S. 108–129.

Swarts, Jason (2009): The collaborative construction of "fact" on Wikipedia. Proceedings of the 27th ACM international conference on Design of communication. New York, NY, USA: ACM (SIGDOC ’09), S. 281- 288.

Wikipedia, Die freie Enzyklopädie (Hg.) (2013a): Wikipedia:Über Wikipedia. Retrieved March 31, 2013 from http://de.wikipedia.org/w/index.php?title=Wikipedia:%C3%9Cber_Wikipedia&oldid=116077822.

Wikipedia, Die freie Enzyklopädie (Hg.) (2013b): Seiten mit den meisten Versionen. Retrieved March 31, 2013 from http://de.wikipedia.org/wiki/Spezial:Meistbearbeitete_Seiten

Wikipedia, Die freie Enzyklopädie (Hg.) (2013c): Wikipedia: Bewertungen. Retrieved March 31, 2013 from http://de.wikipedia.org/w/index.php?title=Wikipedia:Bewertungen&oldid=115305426

Wikipedia, Die freie Enzyklopädie (Hg.) (2013d): Hilfe:Versionen. Retrieved March 31, 2013 from

https://de.wikipedia.org/w/index.php?title=Hilfe:Versionen&oldid=114258827

Wikipedia, Die freie Enzyklopädie (Hg.) (2013e): Wikipedia:Exzellente Artikel. Retrieved April 15, 2013 from

http://de.wikipedia.org/w/index.php?title=Wikipedia:Exzellente_Artikel&oldid=116699124

Wikipedia, Die freie Enzyklopädie (Hg.) (2013f): Wikipedia: Lesenswerte Artikel. Retrieved April 15, 2013 from

http://de.wikipedia.org/w/index.php?title=Wikipedia:Lesenswerte_Artikel&oldid=116938655

Yaari, Eti; Baruchson-Arbib, Shifra & Bar-Ilan, Judit (2011): Information quality assessment of community generated content: A user study of Wikipedia. In: Journal of Information Science, 37(5):487-498.

Das Spiel dauert 140 Zeichen

Wie deutsche Fußballvereine Twitter für Marketing und Fan-Kommunikation entdecken

Katrin Weller[1], Axel Bruns[2]

[1] GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8
D-50667 Köln
katrin.weller@gesis.org

[2]Queensland University of Technology
Z1-515, Creative Industries Precinct
Musk Ave, Kelvin Grove Qld. 4059, Australien
a.bruns@qut.edu.au

Zusammenfassung

Wie nutzen Vereine der 1. Fußball-Bundesliga das Potential von Twitter für die Marken- und Fan-Kommunikation? Während der Fußballsaison 2011/12 wurden für die Untersuchung einerseits Follower-Zahlen erhoben und andererseits Tweets von Bundesligaclubs gesammelt. Das so entstandene Tweet-Korpus ermöglicht bereits mit der Hilfe von Standard-Metriken einen Überblick über die Nutzungsstrategien. Einzelne Vereine werden als Fallstudien detaillierter betrachtet.

Abstract

This paper investigates how soccer clubs from Germany's first division, the 1. Bundesliga, make use of Twitter for marketing and fan communication. During the 2011/12 season we determined follower numbers and collected tweets from the clubs' Twitter accounts. The resulting corpus can be analysed with standard metrics for general insights. These will be enriched with a closer look at certain clubs.

Twitter und Sportkommunikation

Twitter nimmt für die Kommunikation rund um verschiedene Sportereignisse, etwa Tour de France, Fußball-Europameisterschaft, Super Bowl oder die Olympischen Sommerspiele, eine zunehmend bedeutendere Rolle ein. Twitter selbst versucht seine Position in der Sportberichterstattung stellenweise zu verfestigen, beispielsweise durch die Veröffentlichung spezieller Hashtags, die sich auf einzelne Spieltage und Partien der deutschen Fußball-Bundesliga beziehen (Twitter auf Deutsch, 2012). Zeitweise blendete Twitter auch den aktuellen Spielstand ein, wenn man während eines Livespiels nach einem Vereinsnamen bei Twitter gesucht hat. Neben einer Nachrichtenfunktion kann dabei Twitter auch für die Formierung von Fankulturen eine Rolle zukommen. Auf Twitter findet man sowohl einzelne Sportler (z.B. Radfahrer Lance Armstrong, Tennisspielerin Venus Williams), als auch Vereine, Verbände und Sportjournalisten – und, nicht zuletzt, zahlreiche Fans und Fangruppierungen. Es vermischen sich klassische Berichterstattung mit Hintergrundberichten und privaten Sichtweisen der Sportler (Frederick et al., 2012). Durch die offene Kommunikationsstruktur bietet die Plattform zumindest theoretisch die Chance, einen direkten Austausch zwischen Fans und Sportlern oder Fans und Vereinen zu ermöglichen.

Studien zeigen, dass Online-Räume bei der Identitätsstiftung (Gibbons & Dixon, 2010) und der Community-Bildung (Krøvel, 2012) für Sportfans eine Rolle spielen und dass sie für bestimmte Subkulturen förderlich sind (Merkel, 2012). McLean & Wainwright (2009) sehen im Internet allgemein einen Raum, in dem Fans ihre Unzufriedenheit mit der Vereinsführung ausdrücken und sich gar gegen die kommerzielle Macht der Vereine zur Wehr setzen können. Auch Auty (2002), die die Rolle von Internetangeboten für die Fankultur (englischer) Fußballvereine untersucht, betont, dass außerhalb offizieller Marketingkanäle der Vereine im Internet ein Raum für die Interessensvertretung von Fan-Communities entsteht. Als Beispiel wird der Widerstand gegen den Umzug des Wimbledon Football Club nach Milton Keynes hervorgehoben (Auty, 2002). Insgesamt werden oft zwei verschiedene Motivationslagen und damit in der Forschung zwei verschiedene Untersuchungsbereiche wahrgenommen: die Seite der Vereine, die um ihre öffentliche Darstellung und Marketingmaßnahmen bemüht sind, und die Seite der Fans, die ein Mitspracherecht ausüben wollen. Es gibt jedoch auch Fälle, in denen Online-Räume von Clubs gezielt genutzt werden, um Fans in

Entscheidungen mit einzubeziehen, z.B. beim Leyton Orient Football Club (Auty, 2002) oder bei Ansätzen des Crowdsourcings im Profifußball (Breuer & Trixa, 2012).

Im Zusammenhang mit Twitter wurden mittlerweile verschiedene Aspekte der Sportkommunikation untersucht. In vielen davon steht die Twitter-Nutzung durch individuelle Sportler im Vordergrund; entweder allgemein oder fokussiert auf bestimmte Sportarten (beispielsweise Radfahren, Kassing & Sanderson, 2010) oder besondere Personengruppen (Clavio & Kian, 2010). Frederick et al. (2012) werteten Tweets von Sportlern aus verschiedenen Disziplinen aus und zeigen, dass diese Twitter nutzen, um Fans einen Einblick in ihr Privatleben zu ermöglichen und um mit anderen Nutzern zu interagieren. Hambrick et al. (2010) teilten Sportler-Tweets manuell in sechs verschiedene Kategorien ein: Interactivity, Diversion, Information sharing, Content, Promotional und Fanship. Sowohl hier als auch bei Pagaro (2010) zeigt sich ein hohes Maß an Interaktion zwischen Sportlern und Fans auf Twitter. Vergleichbare Untersuchungen auf Vereinsebene gibt es bislang nicht. Sherwood und Nicholson (2012) fanden durch qualitative Interviews mit australischen Sportjournalisten heraus, dass für diese Twitter die wichtigste Web 2.0-Quelle für ihre Arbeit darstellt.

Fußball ist in Deutschland die populärste Sportart, was sich einer hohen Medienabdeckung widerspiegelt. Die Fußball-Bundesliga ist professionell organisiert und wird intensiv vermarktet; die Vereine der 1. Bundesliga erzielten in 2009/10 gemeinsam rund 1,6 Milliarden Euro an Einnahmen und waren so nach der English Premier League die finanziell zweitstärkste Fußball-Liga in Europa (Deloitte, 2011), gleichzeitig kann die Bundesliga dabei seit Jahren eine Gewinnsteigerung verzeichnen (Bundesliga Report, 2013). Bei den Stadionbesuchern liegt Deutschland europaweit vorne (Deloitte, 2011). Es ist also zu erwarten, dass sich die Bundesligavereine auch in verschiedenen Online-Umgebungen professionell vermarkten. Kernfragestellung dieser Arbeit ist, wie von offizieller Seite im deutschen Profifußball mit den speziellen Möglichkeiten des Kommunikationsraums auf Twitter umgegangen wird: Nutzen Vereine Twitter für einen Diskurs mit ihren Fans? Wird mit der Community interagiert? Wie unterscheiden sich die Herangehensweisen einzelner Vereine? Diesen Fragen wollen wir uns hier annähern, wenngleich sie ohne zusätzliche Arbeiten noch nicht vollständig beantwortet werden können. Dieses Paper nimmt zunächst eine Bestandsaufnahme zur Twitter-Landschaft der 1. Fußball Bundesliga vor und zeigt, wie sich die Vereine im Laufe der Saison 2011/12 bei Twitter etabliert haben. Zudem werden ver-

schiedene Herangehensweisen der Vereine für den Umgang mit Twitter in der Fan-Kommunikation identifiziert. Eine Besonderheit dieser Studie liegt in der Beobachtung einer kompletten Liga. Die meisten anderen Studien im Bereich Sportkommunikation bei Twitter betrachten Stichproben mit verschiedenen Athleten. Darüber hinaus wurde der Beobachtungszeitraum bewusst lang gewählt, betrachtet wird in dieser Arbeit zunächst eine komplette Saison (2011/12); die Datenerhebung läuft über die Saison hinaus bis jetzt weiter, so dass auch noch längere Ausschnitte betrachtet werden können. Diese Arbeit ist Teil eines laufenden Projektes, das die Twitter-Nutzung deutscher, englischer und australischer Fußballvereine vergleicht (Bruns, Weller & Harrington, im Druck). Andere Studien, die verschiedene internationale Ligen miteinander vergleichen, gibt es bislang nicht. Insgesamt soll mit dem Projekt ein Beitrag zur Erforschung von Sportmarketing in Online-Räumen geleistet werden, der dabei hilft, die Lücke zwischen den Forschungsbereichen zu Unternehmenskommunikation und Marketing im Web 2.0 einerseits und der Forschung zu Fankulturen andererseits zu schließen.

Methode

Um ein umfassendes Bild der 1. Fußball-Bundesliga auf Twitter zeichnen zu können, wurden zwei unterschiedliche Arten von Daten gesammelt: Tweets und Follower-Zahlen. Um beides umsetzen zu können, musste in einem ersten Schritt damit begonnen werden, die Twitter-Accounts aller Vereine, die in der Saison 2011/2012 in der 1. Liga vertreten waren, zu identifizieren. Da zu dem Zeitpunkt keine offizielle Übersicht aller Vereins-Accounts verfügbar war[1], wurde manuell bei Twitter nach den Vereinsnamen gesucht. Im Juni 2011, als die Liste der Vereine erstellt wurde, verfügten nur zwei deutsche Erstligavereine über „verifizierte“ Twitter-Konten, hatten sich also von Twitter bescheinigen lassen, dass es sich um offizielle Konten der angegebenen

1 Eine Liste mit Sport-bezogenen Twitter-Accounts, sortiert nach Sportarten, gibt es unter http://www.sportsin140.com/. Sie wird u.a. von Frederick et al. (2012) zur Stichprobenziehung verwendet. Für die Saison 2012/13 hat Twitter auf Deutsch (2012) eine Liste von Bundesliga-Club-Accounts veröffentlicht.

Vereine handelt. Für einige Vereinsnamen wurden mehrere Twitter-Accounts gefunden, bei denen nicht klar war, ob es sich um offizielle Aktivitäten der Vereine oder um von Dritten betriebene Konten handelt (beispielsweise @VfB_de und @VfB für den VfB Stuttgart oder @HSV und @HSV1919 für den Hamburger SV). In solchen Fällen wurde zunächst überprüft, ob von der offiziellen Vereins-Webseite auf einen bestimmten Twitter-Account verwiesen wird. War das der Fall, so wurde der entsprechende Account ausgewählt, im Zweifelsfall wurden mehrere in Frage kommende Accounts weiter beobachtet. In Einzelfällen wurde per E-Mail bei der Pressestelle der Vereine nachgefragt, insbesondere beim FC Bayern München, für den kein plausibler Account gefunden werden konnte (s.u.). Es entstand so im Juni 2011 eine Liste von 25 Twitter-Konten, die fortan beobachtet werden sollten. Für diese Nutzernamen wurden (1) alle vom den Accounts verfassten Tweets und alle an den Account adressierten Tweets (@-Mentions/Retweets) für die gesamte Saison 2011/12 gesammelt und (2) regelmäßig einmal pro Monat die Anzahl der Follower erhoben. Die Erhebung der Follower-Zahlen wurde manuell durchgeführt, i.d.R. jeweils am dritten Donnerstag eines Monats. In einigen Fällen kam es in der Zwischenzeit zu Namensänderungen oder es wurden neue offizielle Accounts aufgesetzt. Für diese Nutzernamen, die erst im Laufe der Zeit hinzukamen, konnten daher die Tweets erst ab einem späteren Termin im Laufe der Saison erhoben werden. Bei der monatlichen Durchsicht der Accounts zur Erhebung der Follower-Zahlen wurden zudem weitere Besonderheiten vermerkt, beispielsweise, ob ein Account inzwischen verifiziert wurde oder ob im Profil eine besondere Beschreibung auftaucht. Die Archivierung der Tweets erfolgte mit Hilfe der Software YourTwapperkeeper (Bruns, 2011). Bei der Datensammlung unterliegt man den technischen Begrenzungen durch die Twitter-APIs (Gaffney & Puschmann, im Druck; Morstatter et al., 2013), die unter anderem die abrufbaren Tweets auf 1% des Twitter-Gesamtvolumens begrenzen. Zudem kann es durch Serverausfälle o.ä. vorübergehend zu kurzfristigen Datenverlusten kommen. Die Datensammlung basiert auf den ausgewählten Nutzernamen. Ein alternativer Ansatz, bei dem Tweets anhand von Hashtags gesammelt werden können, wurde für dieses Projekt nicht verfolgt. Dies hat Auswirkungen darauf, welche Art von Aussagen auf Grund der Daten getroffen werden können. Die hier erhobenen Daten zeigen, inwiefern die Vereine selbst bei Twitter aktiv sind. Da zusätzlich Tweets gesammelt wurden, die einen Vereins-Account erwähnen (in denen also der Vereins-Nutzername mit vorangestelltem „@“-Symbol vorkommt), lässt sich auch untersuchen, ob andere Twitter-Nutzer

direkt mit den Vereins-Accounts in Verbindung zu treten versuchen. Es kann jedoch keine Aussage dazu getroffen werden, in welchem Ausmaße insgesamt über einzelne Vereine bei Twitter gesprochen wird. Hierzu hätten auch Tweets nach Keywords und Hashtags gesammelt werden müssen. Die gesammelten Tweets wurden nach statistischen Kriterien ausgewertet, wobei auf sich derzeit etablierende Standard-Metriken (Bruns & Stieglitz, 2013) zurückgegriffen werden konnte. Hierzu gehört insbesondere die Auswertung der Tweets pro User unterteilt in Originaltweets, bearbeitete sowie unveränderte Retweets („edited and unedited retweets“) und echte @-Mentions (Bruns & Stieglitz, 2013). Echte @-Mentions („genuine @mentions“) sind Tweets, die einen Nutzernamen in der Form @username enthalten, aber keine Retweets sind. Alle diese Tweet-Arten können zudem URLs enthalten, der Anteil der Tweets mit URL wird daher separat erfasst. Diese Metriken helfen, die Aktivität einzelner User zu charakterisieren und zu vergleichen. Zusätzlich ist es sinnvoll, die Sichtbarkeit („visibility“) eines Accounts zu untersuchen, indem man die Retweets und @-Mentions untersucht, die dieser Account erhält. Eine komplette inhaltliche Analyse der Tweets ist bei der gegebenen Datenmenge nicht möglich. Eine spezialisierte Inhaltsanalyse für unsere Ausgangsdaten ist für weiterführende Studien geplant; hier wurde in den Fällen, in denen sich besondere Auffälligkeiten ergaben, auf der inhaltlichen Ebene die Tweets stichprobenartig überprüft.

Erste Ergebnisse zur Saison 2011/2012

Die Bundesliga etabliert sich auf Twitter

Was den Einstieg bei Twitter angeht, so präsentierten sich die Erstligaclubs zunächst wenig professionell. Das zeigt sich am deutlichsten daran, dass im Juni 2011 noch kaum verifizierte, offizielle Twitter-Accounts üblich sind. Im Laufe der Saison 2011/12 konnte jedoch eine maßgebliche Entwicklung der Vereins-Accounts beobachtet werden, auch einhergehend mit zahlreichen

Namenswechseln[2]. Die Anzahl der Follower stieg dabei von Juni 2011 bis Juni 2012 stetig (Abbildung 1), je nach Verein um das 1,05-Fache bis 1,33-Fache pro Monat. Diesen Trend unterbrechen konnten lediglich die zwischenzeitliche Auflösung oder Umbenennung einzelner Accounts.

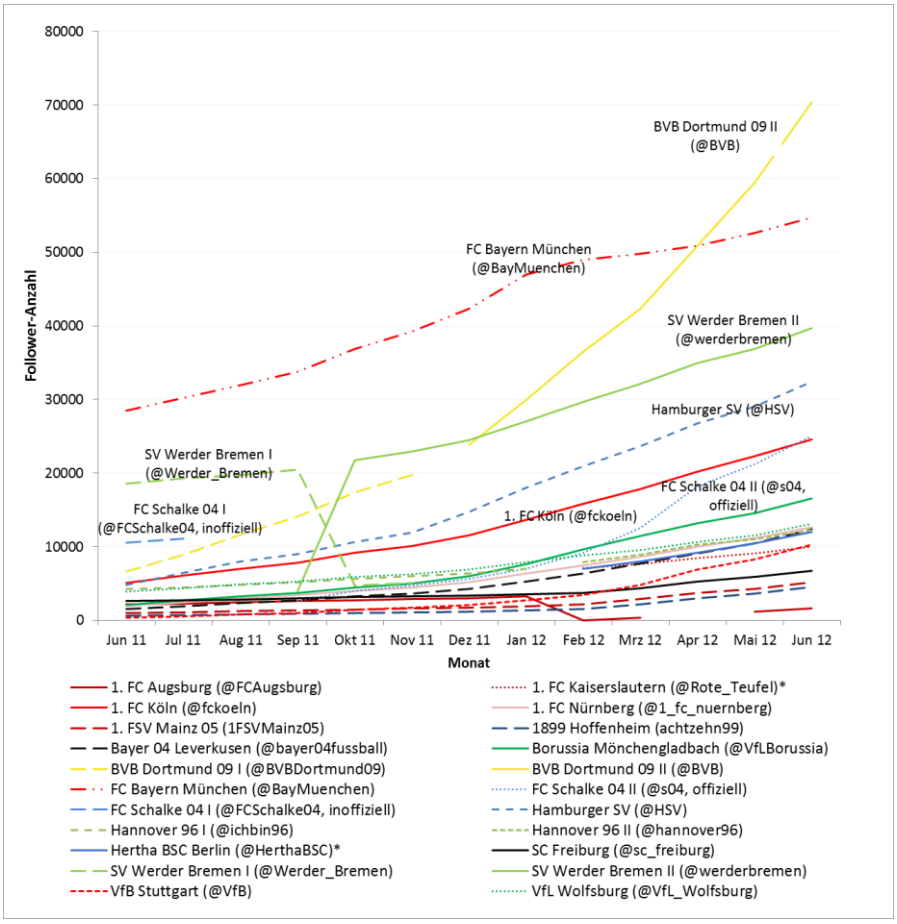

Abbildung 1. Entwicklung der Follower-Zahlen während der Saison 2011/12.

2 Zum Vergleich: in der English Premier League gab es im Juni 2011 fünf Vereine mit verifiziertem Account, darunter vier der fünf großen Vereine Arsenal, Liverpool, Chelsea und Manchester City. Die weiteren Clubs folgten im Laufe der Saison, mit Ausnahme von Manchester United. Namenswechsel kamen hier deutlich seltener vor.

Im Juni 2011 hatten allein der 1. FC Köln sowie der damalige deutsche Meister Borussia Dortmund bereits ihren offiziellen Twitter-Account verifizieren lassen. Ab September folgten Schritt für Schritt auch die anderen Vereine diesem Beispiel. Im März 2012 waren fast alle Erstligavereine mit einem verifizierten Account vertreten; einzig der Rekordmeister Bayern München fehlte. Hier war der Account @BayMuenchen (Abb. 1) zwar zu Testzwecken vom Verein angelegt worden, wurde jedoch nicht weiter genutzt. Erst zu Beginn der Saison 2012/13 zog Bayern mit einem neuen, offiziellen Account (@fcbayern) nach. Dies weist auf ein weiteres Phänomenen hin, dass deutlich macht, dass einige Bundesliga-Vereine nicht von Anfang an eine konsequente Twitter-Strategie entwickelt hatten und sich hier erst professionalisieren mussten: Die Hälfte aller Vereine wechselte seit Juni 2011 ihren Twitter-Namen, erstellte einen neuen Account oder übernahm ein bereits bestehendes Twitter-Konto. Es wurden sperrige Namen verkürzt (z.B. wurde aus @BVBDortmund09 schlicht @BVB, aus @VfL_Borussia wurde @Borussia), manchmal eher spielerische Namen zurückgenommen (aus @ichbin96 wurde @Hannover96). Einzelne Twitter-Konten, deren Namen Variationen von Vereinsnamen enthielten, wurden im Laufe der Saison gesperrt (z.B. @VfB_de, @HSV1919) – ob dies auf Veranlassung der Vereine geschah, lässt sich jedoch meist nicht sagen.

Durch die relativ späte professionelle Nutzung sahen sich manche Vereine auf Twitter mit Accounts konfrontiert, die (von anderen Nutzern betrieben) bereits Nachrichten über den eigenen Verein verbreiteten. Im Falle des 1. FC Schalke 04 war beispielsweise der vom Verein gewünschte Nutzername @s04 bereits vergeben. Unter diesem Namen war ein belgischer Schalke-Fan registriert. In seinem Blog berichtet dieser Fan, wie ihm der Account im August 2011 entzogen wurde (Nettooor, 2011). Schalke nutzte hier nicht die Gelegenheit, sich friedlich mit dem Fan zu einigen, der nach eigener Aussage dem Verein nur helfen wollte: „As I have been a Schalke fan for more than 35 years, it struck me that many big Bundesliga clubs were present on Twitter, but not my Schalke04?“ (Nettooor, 2011). Mehr noch, Schalke verpasste auch die Gelegenheit, bereits bestehende Follower für ihren eigenen Account zu gewinnen. So existierte bis Juli 2011 auch ein weiterer inoffizieller Account, @FCSchalke04, der mit mehr als 10.000 Followern damals der drittbeliebteste Bundesliga-Twitter-Account war (Abb. 1). Auch dieser wurde zwischen Juli und August 2011 eingestellt. Nachdem Schalke mit dem neu übernommenen @s04-Account selbst bei Twitter aktiv wurde, brauchte der Verein einige Zeit, um sich selbst wieder eine Follower-Basis aufzubauen

(Abb. 1). Anders verlief es beim SV Werder Bremen. Die Bremer setzen ihren eigenen Account @werderbremen zunächst parallel zum bereits bestehenden inoffiziellen Fan-Account @Werder_Bremen auf. Im September 2011 fand sich im Fan-Account ein Hinweis, dass es sich um einen privaten Account handelt ("inoffizieller Account von @leanderwattig", 13. Sept. 2011). Der Account-Inhaber wurde nach eigenen Angaben daraufhin von Werder Bremen freundlich kontaktiert, woraufhin er seinen Account freiwillig und einvernehmlich an den Verein übergeben hat (Persönliche Kommunikation, 15. April 2013). Im Anschluss fand sich eine Zeitlang im Profil von @Werder_Bremen ein Verweis auf den neuen offiziellen Bremer Account @werderbremen ("Der ehemalige Info-Service von Werder Bremen. Verfolge Werder jetzt auf der offiziellen Twitter-Seite", 20. Okt. 2011), bevor dieser schließlich eingestellt wurde. Für Werder hat sich durch diesen kooperativen und fließenden Übergang auch kein Bruch in den Follower-Zahlen ergeben (Abb. 1).

Verschiedene Vereinsstrategien

In Deutschland ist nicht nur der Fußballsport extrem professionalisiert, gleichzeitig verfügt Deutschland über eine sehr ausgeprägte Fußball-Fankultur, in der Fans auch ihre Unzufriedenheit mit der zunehmenden Kommerzialisierung artikulieren (Merkel, 2012). Twitter bietet theoretisch eine zusätzliche Kommunikationsplattform, die Fans für den direkten Austausch mit Vereinen und für die öffentliche Meinungsäußerung nutzen können. Aktuell bildet Twitter jedoch eher die Kluft zwischen Bundesliga-Clubs und Fans ab, als dass es in großem Maße zu einem Austausch zwischen beiden Seiten führt. Viele Vereine nutzen Twitter als ein zusätzliches Tool für klassische Marketing-Aktivitäten. Dies zeigt sich bereits an der Zusammensetzung der Tweets vieler Accounts: Die Twitter-eigenen Möglichkeiten zur direkten Interaktion mit anderen Twitter-Nutzern, Retweets und insbesondere @-Replies, werden von den Vereinen nur sehr sparsam genutzt. Die meisten Tweets sind Originaltweets (Abb. 2), was darauf schließen lässt, dass sich die Vereine eher als Sender und Informationslieferant verstehen. Dies steht in Kontrast zu den Ergebnissen von Hambrick et al. (2010); die dort untersuchten Sportler nutzen Twitter besonders stark zur Interaktion mit Fans und nur wenig zur Informationsverbreitung. Bei einzelnen Accounts fällt der extrem hohe Anteil an Tweets mit URL auf (etwa FC Kaiserslautern und Mainz 05).

Ein Blick in die gesammelten Tweets des Mainzer Accounts bestätigt dies: Eine Überprüfung von 20 zufällig ausgewählten Tweets zeigt, dass von diesen 18 eine URL enthalten, dabei handelt es sich in 17 Fällen um Links auf einen Beitrag innerhalb der Facebook-Seite des Vereins und in einem Fall auf einen Link zur Vereins-Webseite.

Dass auch die Fußball-Fans bei Twitter nicht nur an einem direkten Austausch mit ihren Vereinen interessiert sind, zeigt sich eindrucksvoll am Beispiel von Rekordmeister FC Bayern München. Der Account @BayMuenchen führt lange die Twitter-Follower-Tabelle (Abb. 1) an und wurde erst im Frühjahr 2012 von Borussia Dortmund überholt. Dies ist insofern erstaunlich, als dass @BayMuenchen ein Account ist, der zwar nach Angaben der Münchener Pressestelle einmal zu Testzwecken angelegt wurde, jedoch (fast) keine Tweets veröffentlichte. Die Fans waren hier offenbar damit zufrieden, ihre Verbundenheit mit dem Verein über die Follower-Beziehung auszudrücken oder folgten in der Hoffnung, dass er noch aktiv werden würde.

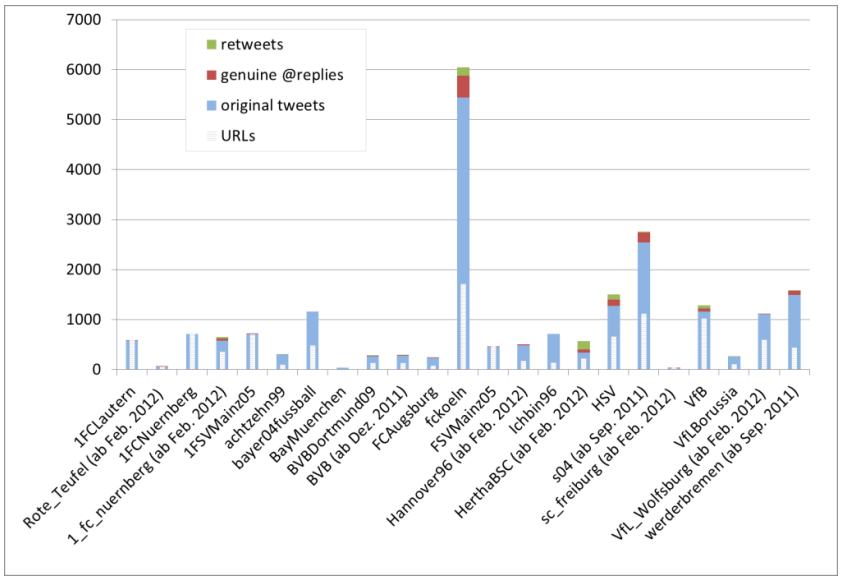

Abbildung 2. Twitter-Aktivitäten der Bundesliga-Clubs, Saison 2011/12.

Der 1. FC Köln ist in anderer Hinsicht etwas Besonderes. Was die Vereinsmitglieder angeht, kann er mit den deutschen Top-Clubs mithalten und liegt dort aktuell mit über 50.000 Fans auf Platz 6. In der untersuchten Saison

2011/12 spielte Köln jedoch so mäßig, dass es am Ende mit Platz 17 nicht für den Klassenerhalt reichte. Die Followerzahlen auf Twitter stiegen jedoch beständig und Köln belegte auch hier zu Saisonende den 6. Platz. Möglicher Frust über die schlechte Saisonleistung schlägt sich hier somit nicht in Form von Follower-Rückgängen nieder. Führend ist Köln bei der Twitter-Aktivität (Abb. 2) und auch bei den Fan-Reaktionen in Form von Retweets und @-Mentions durch andere Nutzer schneidet der FC quantitativ sehr gut ab. Hier wäre zusätzlich eine detaillierte qualitative Analyse zur Stimmungslage sinnvoll. Stichproben von Tweets am 5. Mai 2012 (als der Abstieg feststand) zeigen sowohl positive („@fckoeln Kopf hoch #effzeh, ihr kommt wieder zurück. Ett kütt wie ett kütt un ett häd imma noch joot jejange.“) als auch negative Reaktionen („@fckoeln Selbst Schuld hättet ihr vielleicht mal früher drüber nachdenken sollen ! Der Vorstand ist einzig und allein ein Karnevalsverein.“). Mehr als andere Bundesliga-Vereine tritt Köln mit anderen Twitter-Nutzern in Kontakt, veröffentlicht Retweets und @-Mentions. Die generell hohe Anzahl erklärt sich einerseits dadurch, dass Twitter zur Live-Berichterstattung während der FC-Spiele eingesetzt wird (die Tweets sind dann mit dem Hashtag #live, manchmal auch mit der Spielminute gekennzeichnet). @fckoeln berichtete aber auch live von einer Präsidentschaftswahl und bezieht Stellung zum Abstiegskampf. Darüber hinaus hat Köln spezielle Twitter-Aktionen umgesetzt, beispielsweise Interview-Sessions, bei denen u.a. der damalige Trainer Solbakken über Twitter auch auf sehr persönliche Fragen der Fans antwortete. Dies ist schon allein deshalb eine Besonderheit, weil nach aktuellem Stand kein einziger Bundesligatrainer selbst einen Twitter-Account betreibt.

Fazit und Ausblick

Im Laufe der Saison 2011/12 haben sich die deutschen Erstligavereine mit eigenen Accounts auf Twitter eingefunden; dabei gab es unterschiedlich professionelle Herangehensweisen. Bislang nutzen viele Vereine Twitter als Marketingkanal eher zur Informationsverbreitung, nur in Ausnahmefällen findet eine direkte Interaktion mit Fans und Followern statt. In Zukunft soll vertiefend untersucht werden, wie sich die Tweet-Inhalte der Vereine unter-

scheiden (in Anlehnung an die Kategorien von Hambrick et al., 2010). Interessant wäre künftig ein Vergleich mit den Facebook-Aktivitäten der Clubs.

Literaturverzeichnis

Auty, C. (2002). Football fan power and the internet: Net gains? Aslib Proceedings, 54(1), 273-279.

Breuer, J., & Trixa, J. (2012). 11.000 Freunde must ihr sein: Fanbeteiligung und Crowdsourcing im Internet am Beispiel von deinfussballclub.de. In C. Brandt, F. Hertel & C. Stassek (Hrsg.), Gesellschaftsspiel Fußball. Eine sozialwissenschaftliche Annäherung (pp. 227-248). Wiesbaden: Springer.

Bruns, A. (2011). Switching from Twapperkeeper to YourTwapperkeeper. Blog post, June 21, 2011, retrieved from: http://mappingonlinepublics.net /2011/06/21/switching-from-twapperkeeper-to-yourtwapperkeeper/

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. International Journal of Social Research Methodology. DOI: 10.1080/13645579.2012.756095

Bruns, A., Weller, K., & Harrington, S. (im Druck). Twitter and Sports: Football Fandom in Emerging and Established Markets. Erscheint in: Twitter and Society. New York: Peter Lang.

Bundesliga Report (2013). Die wirtschaftliche Situation im Lizenzfußball. Retrieved from http://static.bundesliga.de/media/native/autosync/report_2013 _dt_72dpi.pdf

Clavio, G., & Kian, T.M. (2010). Uses and gratifications of a retired female athlete's Twitter followers. International Journal of Sport Communication, 3(4), 485-500.

Deloitte (2011). Annual review of football finance 2011: Highlights. Retrieved from http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local% 20Assets/Documents/Industries/Sports%20Business%20Group/uk_sbg_arff11_hi ghlights.pdf

Frederick, E., Lim, C. H., Clavio, G., Pedersen, P. M., & Burch, L. M. (2012). Choosing between one-way or two-way street: An exploration of relationship promotion by professional athletes on Twitter. Communication & Sport, doi: 10.1177/2167479512466387.

Gaffney, D., & Puschmann, C. (im Druck). Data collection on Twitter. Erscheint in: Twitter and Society. New York: Peter Lang.

Gibbons, T., & Dixon, K. (2010). Surf's Up! A call to take English soccer fan interactions on the Internet more seriously. Soccer and Society, 11(5), 599-613.

Hambrick, M.E., Simmons, J.M., Greenhaigh, G.P., & Greenwell, T.C. (2010). Understanding professional athletes' use of Twitter: A content analysis of athlete tweets. Journal of Sport Communication, 3(4), 454-471.

Kassing, J.W., & Sanderson, J. (2010). Fan-athlete interaction and Twitter: Tweeting through the Giro. A Case Study. International Journal of Sport Communication, 3(1), 113-128.

Krøvel, R. (2012). New media and identity among fans of a Norwegian football club. First Monday, 17(5-7). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2882/3208

McLean, R., & Wainwright, D.W. (2009). Social networks, football fans, fantasy and reality: How corporate and media interests are invading our lifeworld. Journal of Information, Communication and Ethics in Society, 7(1), 54-71.

Merkel, U. (2012). Football fans and clubs in Germany: Conflicts, crises and compromises. Soccer and Society, 13(3), 359-376.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In The 7th International Conference on Weblogs and Social Media (ICWSM-13), Boston, MA. Retrieved from http://www.public.asu.edu/~fmorstat/paperpdfs/icwsm2013.pdf

Nettooor (2011). Account @S04 suspended #fail Schalke. 30 Aug. 2011. Retrieved from http://www.nettooor.be/?p=1327

Pagaro, A. (2010). Look who's talking – Athletes on Twitter: A case study. International Journal of Sport Communication, 3(4), 501-514.

Scherwood, M., & Nicholson, M. (2012). Web 2.0 platforms and the work of newspaper sport journalists.

Twitter auf Deutsch (2012). Die Fußball-Saison 2012/2013 auf Twitter. 24 Aug. 2012. Retrieved from http://blog.de.twitter.com/2012/08/die-fuball-saison-20122013-auf-twitter.html