Internet and Web Algorithms – Final Project - Dr. Fred Annexstein

Blaine Booher

June 10, 2008

As my research topic I chose five research papers that have the common theme of working with Wikipedia. Wikipedia is the online encyclopedia where content is driven by open community support and contribution. The encyclopedia's platform is revolutionary because it allows anyone to add information, without requiring authentication. The first paper topic deals directly with Wikipedia. The research performed focuses on what exactly drives the users and contributors, and creates a model to predict useful behavior. The second paper is more technical, and also deals directly with Wikipedia as the focus. The paper creates a model for accurately detecting vandalism in Wikipedia articles. The third paper focuses on how to track and benchmark the evolution of schema in Wikipedia releases. The authors present a benchmark suite for allowing tracking data base structural changes and their impact on existing software. It is a very interesting look into how one can manage a software project to allow for massive expansion, while minimizing the interoperability impact on third party software using a standard interface. The fourth and fifth papers are related in that they use Wikipedia as a guide for formulating text-based categorization models.

The first paper is a summary of research done by a group at the University of Wuerzburg. The title of the paper is Voluntary Engagement in an Open Web-based Encyclopedia: From Reading to Contributing (Schroer and Hertel). The goal of the paper is to find some useful metrics for figuring out what drives contributors and readers in their usages and contributions to Wikipedia. The motivation derives from the fact that a crucial element of Wikipedia success is its ability to attract new users who are also able to create new content and/or enhance existing content. The first measurement was on motivational determinants of social movement participation. The second measurement looked at highly active Wikipedia contributors. The paper creates a prediction model for **willingness to contribute** and a**ctual engagement**. There were two stages of research performed – an initial inquiry done on participants, and a follow up six months later to measure continued participation. The following table shows the highest motivators for reasons people **did contribute** and **did not contribute**, respectively. This paper is not as technical as the other papers, however I did find this to be one of the most interesting studies. I am always very curious as to what motivates people to contribute content to a service in which they are not financially compensated. The study shows that perceived costs and benefits, expected task enjoyment, and the perceived instrumentality of one's own contribution are the strongest factors in determining the willingness of someone to be a contributor. Among other things, actual contribution is triggered by inaccurate or missing articles especially if you are an expert in your field.

Ranking of Motives: Pros and Cons

Readers who contributed to Wikipedia between T1 und T2 ranked the importance of several motives for their engagement ($n = 14$; 18 motives offered, 1 = most important, 18 = least important).

Rank	Motive
3.71	Free access to knowledge for everyone
5.15	Task enjoyment / Fun
5.33	Learning
6.55	Belief in the future of Wikipedia
6.69	Existing information was inaccurate
7.25	Quality improvement of Wikipedia
…	…

Similarly, readers who did not contribute to Wikipedia ranked the importance of motives for not contributing ($n = 69$; 18 motives offered).

Rank	Motive
3.59	No time
4.93	Saw no reason to contribute
5.10	Didn't know what to contribute
5.79	Didn't know how / where to start
6.29	Own contribution not important / useful
6.43	Already enough authors
…	…

The second paper focuses on creating a model to accurately and autonomously discover vandalism in Wikipedia articles. The paper is titled Automatic Vandalism Detection in Wikipedia (Potthast, Stein, Gerling). Wikipedia

revolves around the model that anyone can edit and add content, unauthenticated. The problem, however, is that the freedom of editing has been misused by some editors. The paper distinguishes the editors into three groups: 1) Lobbyists who try and push their own agenda, 2) spammers, who solicit products or services, and 3) vandals, who deliberately destroy the work of others. The paper makes mention of other tools that are used by the Wikipedia community such as WikiScanner. Wikiscanner uses geographical and IP information to identify lobbyists from corporations and countries that may be writing biased information. The paper sets up four stages of the research: defining the task as a classification problem, discussing characteristics that humans use to recognize vandalism, developing tailored features to identify them, and creating a machine-readable corpus of vandalism edits as a baseline for future research. The following tables show the statistics for the distribution of vandalism medium per type (See next page). There are several classifiers that indicate the type of vandalism. The vandalism edit is said to have 'point of view' characteristic if personal opinion is expressed. This can be detected by the use of personal pronouns. Many vandalism edits introduce off topic text with respect to surrounding text, are nonsense or contradictory to common sense, or do not form correct sentence structure. Some edits are vulgar, but we do not want to exclude vulgarity in context. A majority of edits are just gobblygook, or random text, while some edits are simple deletions of entire bodies of text. In reality only 5% of edits at any given time are vandalism. The vandalism corpus is made up of known Wikipedia vandalism attempts. Using a logistic regression classifier the researchers were able to evaluate the discriminative power of vandalism classes defined in Table 3. Their results were impressive, with the model developed correctly identified vandalism much more often than the existing technologies. Also the researchers were able to publish their vandalism corpus for further research.

Table 2. Organization of vandalism edits along the dimensions "Edited content" and "Editing category": the matrix shows for each combination the portion of specific vandalism edits at all vandalism edits. For vandalized structure insertion edits and content insertion edits also a list of their typical characteristics is given. It includes both the characteristics described in the previous research and the Wikipedia policies.

Editing category	**Edited content**			
	Text	Structure	Link	Media
Insertion	43.9% Characteristics: point of view, off topic, nonsense, vulgarism, duplication, gobbledegook	14.6% Characteristics: formatting, highlighting	6.9%	0.7%
Replacement	45.8%	15.5%	4.7%	2.0%
Deletion	31.6%	20.3%	22.9%	19.4%

Vandalism Indicating Features. We have manually analyzed 301 cases of vandalism to learn about their characteristics and, based on these insights, to develop a feature set $\mathcal{F}$. Table 2 organizes our findings as a matrix of vandalism edits along the dimensions "Edited content" and "Editing category"; Table 3 summarizes our features.

Table 3. Features which quantify the characteristics of vandalism in Wikipedia

Feature f	**Description**
char distribution	deviation of the edit's character distribution from the expectation
char sequence	longest consecutive sequence of the same character in an edit
compressibility	compression rate of an edit's text
upper case ratio	ratio of upper case letters to all letters of an edit's text
term frequency	average relative frequency of an edit's words in the new revision
longest word	length of the longest word
pronoun frequency	number of pronouns relative to the number of an edit's words (only first-person and second-person pronouns are considered)
pronoun impact	percentage by which an edit's pronouns increase the number of pronouns in the new revision
vulgarism frequency	number of vulgar words relative to the number of an edit's words
vulgarism impact	percentage by which an edit's vulgar words increase the number of vulgar words in the new revision
size ratio	the size of the new version compared to the size of the old one
replacement similarity	similarity of deleted text to the text inserted in exchange
context relation	similarity of the new version to Wikipedia articles found for keywords extracted from the inserted text
anonymity	whether an edit was submitted anonymously, or not
comment length	the character length of the comment supplied with an edit
edits per user	number of previously submitted edits from the same editor or IP

The third paper focused on following the impact of schema evolution on the platform that Wikipedia is based on, Mediawiki. The paper is titled Schema Evolution in Wikipedia: Toward a Web Information System Benchmark (Curino, Moon, Tanca, Zaniolo). This paper focuses on determining the impact that database restructuring between revisions has on third party applications and other pieces of software using the database engine. When a software project starts out small and grows large, many new features and data structures are added. It becomes beneficial to determine the best way to evolve the queries and database structures to provide minimal impact on re-writing the interfaces. The data management core of an information system is the most critical portion to evolve. The complexity of the database and software maintenance grows with the size and complexity of the entire system. Schema evolution has been extensively studied in traditional systems, so this research focused on the evolution of the database structure. The end goal is to provide a way to measure graceful evolution of schemas. Mediawiki is used for analysis because the user base is very large – used by over 30,000 wikis and supporting over 100,000,000 web pages.

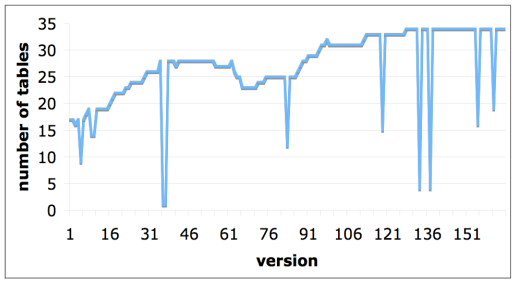


Figure 2: MediaWiki Schema Size: the Number of Tables

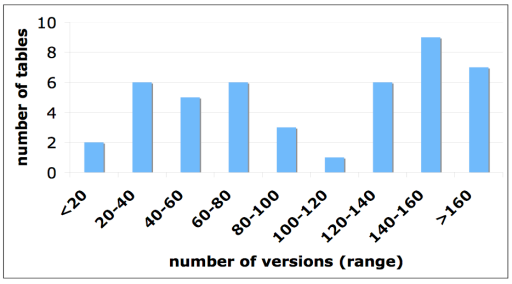


Figure 4: Histogram of Table Lifetime

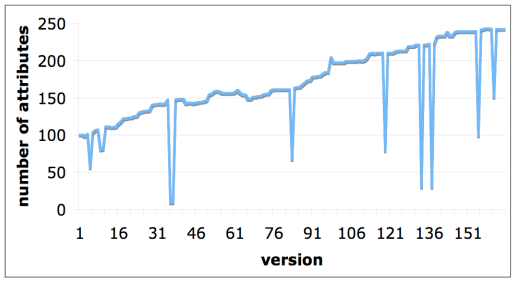


Figure 3: MediaWiki Schema Size: the Total Number of Columns

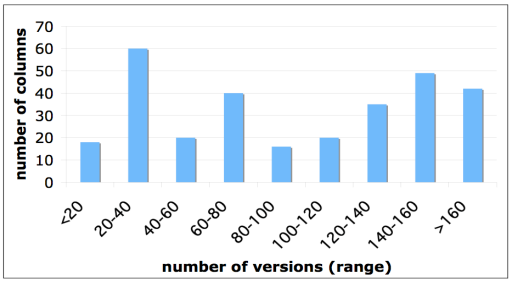


Figure 5: Histogram of Column Lifetime

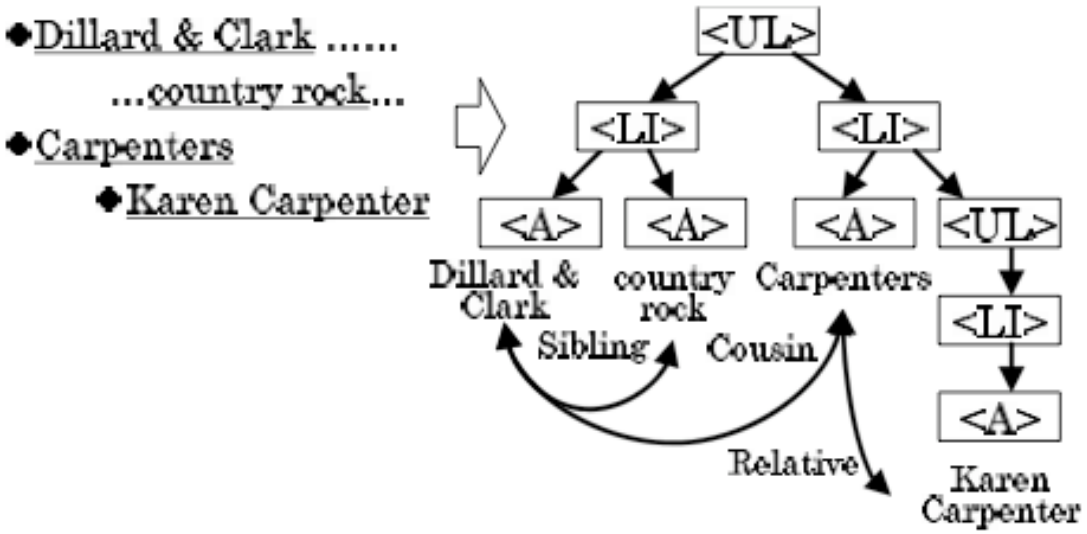
Media wiki has seen 171 different Database schema revisions during its lifetime of almost 5 years. The graph shows a growth trend, where the number of database tables have gone from 17 to 34, and columns from 100 to 242. Sudden drops are due to the revision having syntax errors and breaking the distribution completely

temporarily. Schema growth is due to three main driving forces: performance improvements (caching tables, for example), addition of new features (logging and content validation, for example), and the growing need for preservation of database content history. The researchers are also able to track the life time of certain tables. The core tables, like 'users' have existed since the start of the project, while some tables only stick around for 2 or 3 revisions. In total, 55% of the evolution steps involved actual schema changes, 40% were column or index/key adjustments, 8.8% were rollbacks to previous versions, and 7.6% were only documentation changes. The researchers discovered that at most, the revision evolution has invalidated 32% of the schema interface (existing software calls and queries). Some revisions, under specific circumstances, have invalidated up to 70% of existing queries. The result of the research is a published suite of benchmark tools used to discover metrics and statistics about schema revision evolutions. This is the first time that a unified benchmarking suite has been developed for this purpose. This will help developers track the impact of their changes, and lead to the development of more streamlined schema structures that allow for flexibility while not breaking existing queries.

The fourth paper describes ways that we can use vector space models to develop a kernel for automatic text categorization using the structure of Wikipedia's articles as an existing reference. The paper is titled Wikipedia-based Kernels for Text Categorization (Minier, Bodo, Csato). The research attempts to create intelligent and efficient methods of navigating through a virtual "document space". Document space consists of every document with finite words. Wikipedia contains about 1.6 million articles total. For a more technical insight into the matrices and weight metrics used, reference the included pdf document. The final kernel models a random surfer model using a PageRank derived algorithm. This creates a method of using articles that are more heavily weighted by links that can be used to discover 'document classifications' and relationships. There are also many heuristics and dimension reduction techniques used to more efficiently calculate classification boundaries. A singular value decomposition is used on the matrix tables. The eigenvectors corresponding between two documents give output vectors that give similarity results that can be used to find closeness of classification. The document is trained on a published Reuters corpus from 1987. The results are that the classification of Wikipedia documents has a much higher correlation with human

methods than previous methods. The deviation in results could have been due to the fact that terminology would have changed since 1987, and Wikipedia may be structured differently from how Reuters would have classified their data.

The fifth paper is titled A Graph-based Approach to Named Entity Categorization in Wikipedia using Conditional Random Fields (Watanabe, Asahara, Matsumoto). This paper uses hierarchal data in the XHTML document tree to find 'siblings', 'cousins', and 'children' relationships between concepts. Conditional Random Graphs are used to categorize nodes on the created graph structure. The structures that are most highly used are <UL> or <OL> list objects and <TABLE> objects. The list items, <LI>, tend to be in the same category. The scope of this paper focuses on list items to find category relationships. Anchor texts are also useful in finding dependencies on various categories and improving performance. The HTML document is treated as an ordered tree. The vertices are the anchor texts (<A>) and the edges are cliques of Cousins, Siblings, and Relatives which are derived from the list objects.



In this case, "Dillard & Clark" and "country rock" have a sibling relation, while "Dillard & Clark" and "Carpenters" have a cousin relation since they have a common attribute "Artist". Elements in relation tend to belong to the same class. "Carpenters" and "Karen Carpenter" have a relation in which "Karen Carpenter' is a sibling's grandchild in relation to "Carpenters". In this case the elements tend to be a constituent part of other elements in the relation. The model can capture dependencies by dealing with anchor texts that depend on each other as cliques.

The dataset consists of 2300 articles from the Japanese version of Wikipedia. The researchers use the Extended Named Entity Heirarchy as the class labeling guideline, but reduce the classes to 13 from 200+ by removing similar classes and fine-grained categories. The graph contains Sibling, Cousin, and Relative edges with counts Es=4925, Ec=13134, Er = 746, respectively. (See bottom of this page for reference to results). The results of the experiement are very good. They find that 57% of the NEs can be classified with approximately a 97% accuracy. The NE candidates can be filtered with fewer cost by exploiting marginal probabilities. The researchers mention that they believe they can increase the amount of categorizations by using a more fine grain labeling set to create the categories.

types	feature	SVMs	CRFs
observation features	definition (bag-of-words)	$\surd$	$\surd\,(V)$
	heading of articles	$\surd$	$\surd\,(V)$
	heading of articles (morphemes)	$\surd$	$\surd\,(V)$
	categories articles	$\surd$	$\surd\,(V)$
	categories articles (morphemes)	$\surd$	$\surd\,(V)$
	anchor texts	$\surd$	$\surd\,(V)$
	anchor texts (morphemes)	$\surd$	$\surd\,(V)$
	parent tags of anchor texts	$\surd$	$\surd\,(V)$
	text included in the last header of anchor texts	$\surd$	$\surd\,(V)$
	text included in the last header of anchor texts(morphemes)	$\surd$	$\surd\,(V)$
label features	between-label feature		$\surd\,(S, C, R)$
	previous label	$\surd$	

Table 2: Features used in experiments. "$\surd$" means that the corresponding features are used in classification. The V, S, C and R in CRFs column corresponds to the node, sibling edges, cousin edges and relative edges respectively.

		CRFs								SVMs	
NE CLASS	N	C	CR	I	R	S	SC	SCR	SR	I	P
PERSON	3315	.7419	.7429	.7453	.7458	.7507	.7533	**.7981**	.7515	.7383	.7386
TIMEX/NUMEX	2749	.9936	**.9944**	.9940	.9936	.9938	.9931	.9933	.9940	.9933	.9935
FACILITY	2449	.8546	.8541	.8540	.8516	.8500	.8530	.8495	.8495	.8504	**.8560**
PRODUCT	1664	.7414	**.7540**	.7164	.7208	.7130	.7371	.7418	.7187	.7154	.7135
LOCATION	1480	**.7265**	.7239	.6989	.7048	.6974	.7210	.7232	.7033	.7022	.7132
NATURAL_OBJECTS	1132	.3333	.3422	.3476	.3513	.3547	.3294	.3304	.3316	**.3670**	.3326
ORGANIZATION	991	.7122	.7160	.7100	.7073	.7122	.6961	.5580	.7109	.7141	**.7180**
VOCATION	303	.9088	.9050	.9075	.9059	.9150	.9122	.9100	**.9186**	.9091	.9069
EVENT	121	.2740	.2345	.2533	.2667	.2800	.2740	.2759	.2667	.3418	**.3500**
TITLE	42	.1702	.0889	.2800	.2800	**.3462**	.2083	.1277	**.3462**	.2593	.2642
NAME_OTHER	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	**.0690**	.0000
UNIT	15	.2353	.1250	.2353	.2353	.2353	.1250	.1250	.2353	**.3333**	.3158
ALL	14285	.7846	**.7862**	.7806	.7814	.7817	.7856	.7854	.7823	.7790	.7798
ALL (no articles)	3898	.5476	**.5495**	.5249	.5274	.5272	.5484	.5465	.5224	.5278	.5386

Table 3: Comparison of F1-values of CRFs and SVMs.