

Behavior Mining in Wikipedia

Ee-Peng Lim

Singapore Management University

**The 4th Annual Conference on Collaboration Technologies 2008
Wakayama, Japan, August 30-31, 2008**

Acknowledgement

- Research funding:
 - Agency for Science, Technology and Research (A*Star)
- Collaborators:
 - Hady Wirawan Lauw (Microsoft Research)
 - Aixin Sun (Nanyang Technological University)
- Students:
 - Meiqun Hu (PhD)
 - Ba-Quy Vuong (Undergrad)
 - Minh-Tam Le (Research assistant)

Outline

- Wikipedia: An Overview
- What is behavior mining?
- Article Quality in Wikipedia
- Controversy in Wikipedia
- Conclusion

Web 2.0

- Information source + User participation
 - Network as platform
 - Users jointly own the data
 - Participation
 - Rich, interactive, user-friendly interface
 - Social networking

Wikipedia

- To produce a **free** encyclopedia in each **language** through **collaborative** editing.
- Brief history:
 - English Wikipedia was created in January 2001 by James Wales and Larry Sanger
 - Non-English Wikipedia started to appear in May 2001
 - Managed by Wikimedia Foundation

Wikipedia

WIKIPEDIA

English

2 500 000+ articles

Español

389 000+ artículos

Français

693 000+ articles

Polski

527 000+ haset

Deutsch

790 000+ Artikel

日本語

フリー百科事典

512 000+ 記事

Português

421 000+ artigos

Русский

308 000+ статей

中文

自由的百科全書

203 000+ 條目

Italiano

483 000+ voci



Wikipedia Statistics* – as at last week

- 2,528,073 articles
- 806,030 media files
- 7,709,469 registered users
- 1,585 are admin users

- **Probably the largest collaborative system**

* English version only

Wikipedia Article

[article](#)[discussion](#)[view source](#)[history](#)

2008 Summer Olympics

From Wikipedia, the free encyclopedia

(Redirected from [Beijing Olympics](#))

"Beijing 2008" redirects here. For the video game, see [Beijing 2008 \(video game\)](#).

The **2008 Summer Olympic Games**, officially known as the **Games of the XXIX Olympiad**, was a major international multi-sport event that took place in [Beijing, People's Republic of China](#), from [August 8](#) (except [football](#), which started on [August 6](#)) to [August 24, 2008](#). A total of 10,500 athletes competed in 302 events in 28 sports, one event more than was on the schedule of the [2004 games](#).^[2] The 2008 Beijing Olympics also marked the third time that Olympic events have been held in the territories of two different [National Olympic Committees](#) (NOC), as the equestrian events were being held in [Hong Kong](#).

The Olympic games were awarded to Beijing after an [exhaustive ballot](#) of the [International Olympic Committee](#) (IOC) on [July 13, 2001](#). The official logo of the games, titled "[Dancing Beijing](#)," features a stylised calligraphic character *jīng* (京, meaning *capital*), referring to the host city. The mascots of Beijing 2008 were the five [Fuwa](#),^[3] each representing both a colour of the [Olympic rings](#) and a symbol of Chinese culture. The Olympic slogan, *One World, One Dream*, called upon the world to unite in the Olympic spirit. Several new NOCs have also been recognised by the IOC. The 2008 Olympics was the third time the Olympics had taken place on the Asian continent, and the fifth time for an Olympics outside of Europe and North America.

Games of the XXIX Olympiad



The "[Dancing Beijing](#)" emblem, depicting a [Chinese seal](#) inscribed with the character "Jīng" (京, from the name of the host city) in the form of a dancing figure.

Host city	Beijing, China
Motto	<i>同一个世界 同一个梦想</i> (<i>One World, One Dream</i>)

Revision History

Revision history of 2008 Summer Olympics

From Wikipedia, the free encyclopedia

[View logs for this page](#)

Search in history

From year (and earlier): From month (and earlier):

(Latest | [Earliest](#)) View (newer 50) (older 50) (20 | 50 | 100 | 250 | 500)

For any version listed below, click on its date to view it. For more help, see [Help:Page history](#) and [Help:Edit summary](#).

To search within page histories, try [WikiBlame](#).

(cur) = difference from current version, (last) = difference from preceding version,

m = minor edit, → = section edit, ← = automatic edit summary

Compare selected versions

- (cur) (last) 09:39, 25 August 2008 [Spinner145](#) (Talk | contribs) (65,601 bytes) (→Opening ceremony: remove unverified claim that both girls were listed in credits. If somebody can verify, feel free to put back in.)
- (cur) (last) 08:59, 25 August 2008 [F](#) (Talk | contribs) m (65,666 bytes) (→Opening ceremony)
- (cur) (last) 08:54, 25 August 2008 [Xeltran](#) (Talk | contribs) (65,658 bytes) (→Transport: Changing present/future tense verbs to past; changed "moves" to "transported")
- (cur) (last) 08:48, 25 August 2008 [Xeltran](#) (Talk | contribs) (65,655 bytes) (→Venues: Copyedit; changing present and future tense verbs to past)
- (cur) (last) 07:09, 25 August 2008 [Mervyn](#) (Talk | contribs) (65,698 bytes) (→Closing ceremony: organized by NOT by)
- (cur) (last) 06:40, 25 August 2008 [Yyyzzz](#) (Talk | contribs) m (65,688 bytes) (Chinese officials have not concealed this and later claimed that Li was "a better performer.")
- (cur) (last) 06:06, 25 August 2008 [Xeltran](#) (Talk | contribs) (65,678 bytes) (Changed motto from *Traditional Chinese* to *Simplified Chinese* text, as per talk page discussion)
- (cur) (last) 05:07, 25 August 2008 [DanielEng](#) (Talk | contribs) (65,678 bytes) (→Concerns and controversies)

Wikipedia Article – Wakayama

[article](#)

[discussion](#)

[edit this page](#)

[history](#)

Wakayama, Wakayama

From Wikipedia, the free encyclopedia
(Redirected from [Wakayama](#))

Wakayama (和歌山市 *Wakayama-shi*[?]) is the capital city of [Wakayama Prefecture](#) in the [Kansai](#) region of [Japan](#).

Background

[\[edit\]](#)

It occupies 4 percent of the land area, and has 40 percent of the population, of the prefecture. The city was founded on [April 1, 1889](#).

The city [population](#) rose from 382,155 in 2003 to 386,501 in 2004, a growth of 1.87 percent. The [density](#) as of 2003 was 1,826.74 persons per km². The total area is 209.20 km².

This population increase has occurred despite Wakayama's beleaguered economy, which has suffered since [Sumitomo Steel](#) moved much of its steel producing operations to [China](#). The Wakayama steel mills have since been reduced and restructured, completely shutting in 2004. Additionally, Wakayama is famous across Japan for its [umeboshi](#) and [mikan](#).

Wakayama 和歌山市



Wakayama's location in [Wakayama Prefecture, Japan](#).



Wakayama's location in [Japan](#).

Revision History

(cur) = difference from current version, (last) = difference from preceding version,

m = minor edit, **→** = section edit, **←** = automatic edit summary

Compare selected versions

(cur) (last) 22:19, 16 August 2008 Thijs!bot (Talk | contribs) **m** (3,380 bytes) (*robot Modifying: uk:Вакаяма*) (undo)

- (cur) (last) 19:34, 16 August 2008 Thijs!bot (Talk | contribs) **m** (3,396 bytes) (*robot Adding: uk:Вакаяма, Вакаяма*) (undo)
- (cur) (last) 15:37, 12 August 2008 Kzaral (Talk | contribs) (3,358 bytes) (*Location map*) (undo)
- (cur) (last) 14:18, 31 July 2008 Kzaral (Talk | contribs) (3,277 bytes) (undo)
- (cur) (last) 03:46, 24 June 2008 DumZiBoT (Talk | contribs) **m** (3,203 bytes) (*robot Adding: it:Wakayama*) (undo)
- (cur) (last) 23:38, 28 May 2008 74.186.135.24 (Talk) (3,187 bytes) (undo)
- (cur) (last) 05:47, 24 May 2008 Hmains (Talk | contribs) (3,169 bytes) (*revise category and/or AWB general fixes using AWB*) (undo)
- (cur) (last) 07:51, 23 May 2008 JAnDbot (Talk | contribs) **m** (3,132 bytes) (*robot Adding: tl:Lungsod ng Wakayama*) (undo)
- (cur) (last) 12:17, 17 May 2008 Osm agha (Talk | contribs) **m** (3,105 bytes) (*ar*) (undo)
- (cur) (last) 02:57, 16 May 2008 Fg2 (Talk | contribs) (3,081 bytes) (*Removed Sinaloa. It's a sister of Wakayama Prefecture, not the city of Wakayama. I added it to the article on the prefecture.*) (undo)
- (cur) (last) 02:13, 16 May 2008 201.165.112.194 (Talk) (3,134 bytes) (*→Sister cities*) (undo)
- (cur) (last) 02:12, 16 May 2008 201.165.112.194 (Talk) (3,135 bytes) (*→Sister cities*) (undo)
- (cur) (last) 02:12, 16 May 2008 201.165.112.194 (Talk) (3,137 bytes) (*→Sister cities*) (undo)
- (cur) (last) 08:00, 13 April 2008 Chobot (Talk | contribs) **m** (3,081 bytes) (*robot Modifying: ru:Вакаяма (город)*) (undo)
- (cur) (last) 20:23, 11 April 2008 212.54.28.123 (Talk) (3,084 bytes) (*→Background*) (undo)
- (cur) (last) 14:57, 11 April 2008 72.21.98.60 (Talk) (3,123 bytes) (undo)

(cur) (last) 14:20, 26 March 2008 ClueBot (Talk | contribs) **m** (3,086 bytes) (*Reverting possible vandalism by 86.41.1.1 version by Idioma-bot. False positive? Report it. Thanks, User:ClueBot. (288669) (Bot)*) (undo)

Compare Revisions

[article](#)[discussion](#)[edit this page](#)[history](#)

Wakayama, Wakayama

Coordinates:  34°14′N 135°10′﻿ / ﻿

From Wikipedia, the free encyclopedia

[\(Difference between revisions\)](#)Revision as of 14:20, 26 March 2008 ([edit](#))86.41.129.79 ([Talk](#))Current revision (22:19, 16 August 2008) ([edit](#)) ([undo](#))Thijs!bot ([Talk](#) | [contribs](#))**m** (*robot* *Modifying*: *uk:Вакаяма*)[← Previous edit](#)

(16 intermediate revisions not shown.)

Line 1:`{{Infobox City Japan``|Name= Wakayama``-|JapaneseName= GODZILLA!!!!``|MapImage= Map Wakayama en.png``|Region= [[Kansai]]`**Line 10:**`|Population= 373,655``|Density_km2=``-|Coords= {{coor dm|34|14|N|135|10|E|region:JP_type:city}}``|Tree=``|Flower=`**Line 1:**`{{Infobox City Japan``|Name= Wakayama``+|JapaneseName= 和歌山市``|MapImage= Map Wakayama en.png``|Region= [[Kansai]]`**Line 10:**`|Population= 373,655``|Density_km2=``+|Coords=``+|LatitudeDegrees= 34``+|LatitudeMinutes= 14``+|LatitudeSeconds=``+|LongitudeDegrees= 135``+|LongitudeMinutes= 10``+|LongitudeSeconds=``|Tree=``|Flower=`

What is Behavior Mining?

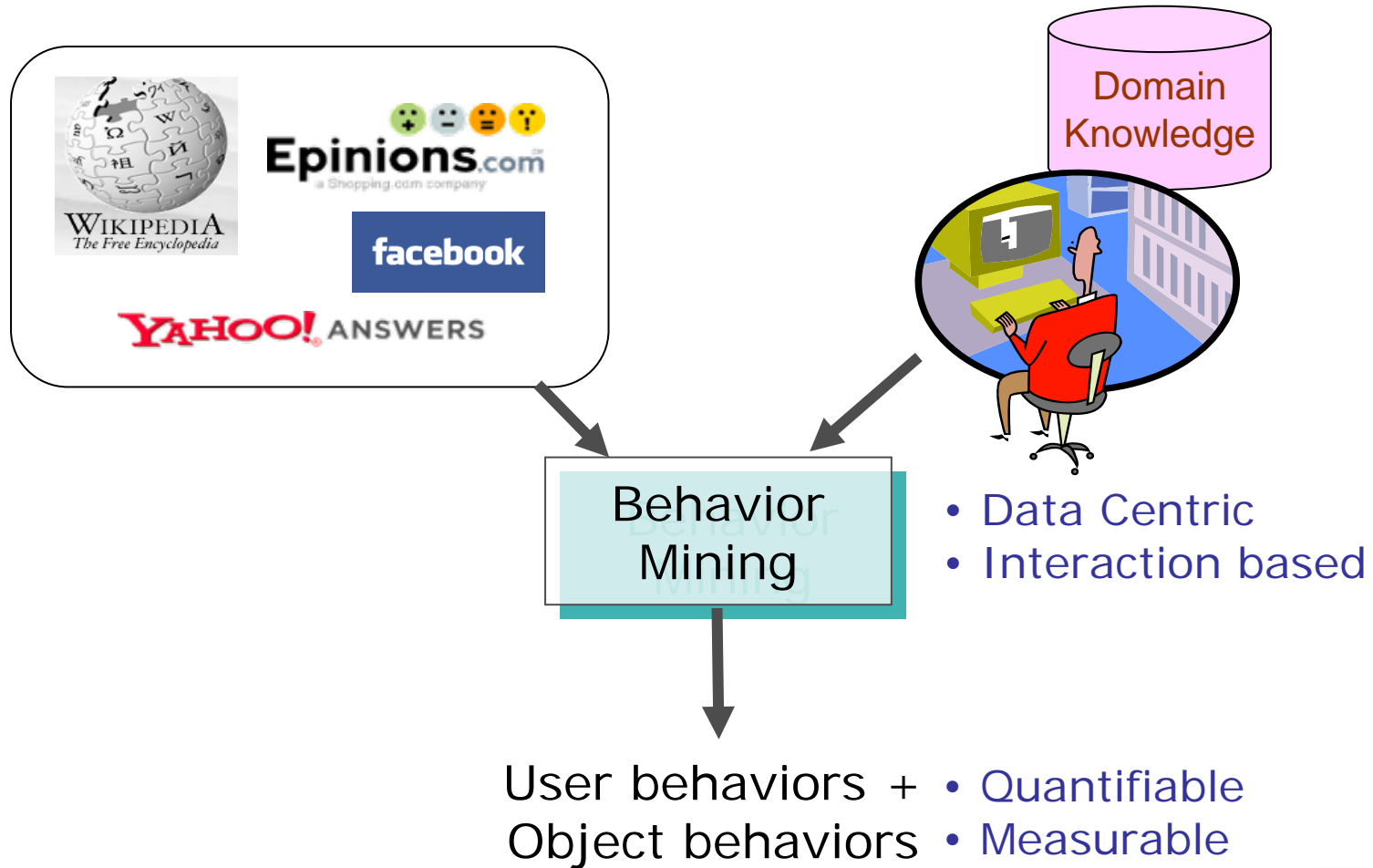
Behaviors

- Behaviors are *patterns* and *properties* of users and objects that affect how they act or be acted upon.
- Examples:
 - Trustworthy
 - Authority
 - Quality
 - Controversy

Behavior Mining

- Determine behaviors from different types of interactions:
 - User-user interactions
 - Users discuss
 - Users give barnstars to other users
 - Administrators block/unblock other users
 - User-object interactions
 - Users create articles
 - Users edit content of articles
 - Administrators protect/unprotect articles

Behavior Mining from User Generation Data



Article Quality in Wikipedia

Article Quality in Wikipedia

- Wikipedia quality is not uniform
 - Middlebury College's history department bans citations of Wikipedia articles (Mar 07)
 - *Nature*: Scientific entries in Wikipedia are of quality comparable to those in Britannica (Dec 05)
- Quality problems
 - Genuine errors
 - [Wikispams](#)
 - Vandalism
 - Hoax, e.g., [John Seigenthaler](#)

Existing Approach in Wikipedia

- Users to determine the good articles – **featured articles**
- **Featured Article (FA):**
 - It is
 - well-written
 - comprehensive
 - factually accurate
 - neutral
 - stable
 - It follows the style guidelines, including the provision of:
 - a lead
 - appropriate structure
 - consistent citations
 - Images
 - Length

Challenges

- Large number of articles
- Wide range of subject topics
- Evolving content in the articles
- Varying contributor background
- Abuses

Rigor and Diversity Models

- Rigor model: More edited articles are good
- Diversity model: Articles edited by many unique users are good

[Lih, Int'l Symposium on Online Journalism 2004]

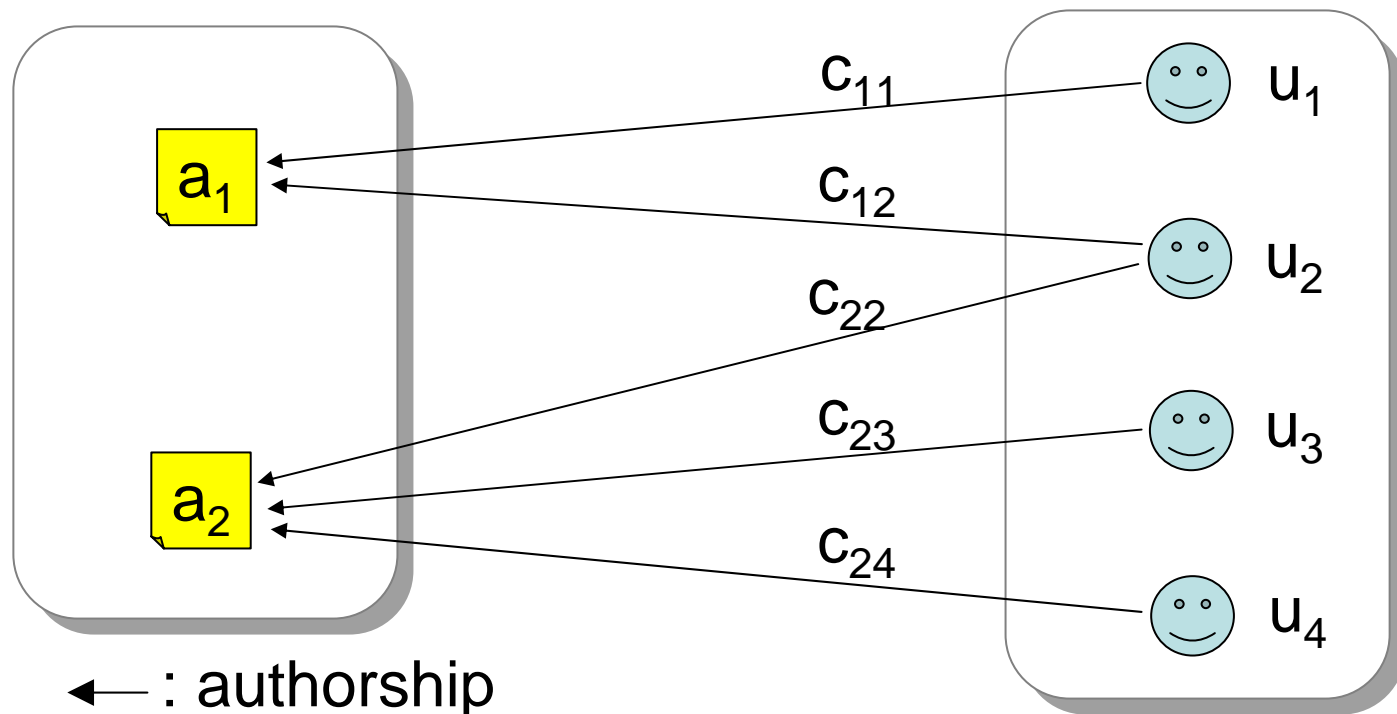
Mining Article Quality

- Can we mine article quality using more robust behavior mining?
- Our approach: **Dependency-aware Behavior Mining**
 - There are often dependencies among behaviors
 - They should be considered in mining behaviors
- Applications:
 - To help users judge the quality of articles
 - To help Wikipedians improve articles
 - To improve Wikipedia search and browsing
- Ideal case:
 - Achieve the above objectives without examining content

Behavior Mining based on Mutual Dependency: Basic Model

Articles (Quality Q_i)

Users (Authority A_j)



Basic Model:

Quality and Authority Behaviors

- An article is an aggregated efforts of multiple authors
- Assumption:
 - Quality:
 - An article has good quality if it is authored by high authority users
 - Authority:
 - A user has high authority if s/he contributes good quality articles

Basic Model

- Quality of article a_i

$$Q_i = \sum_j c_{ij} \cdot A_j$$

- Authority of contributor u_j

$$A_j = \sum_i c_{ij} \cdot Q_i$$

- a_i has words $\{ w_{ik} \}$

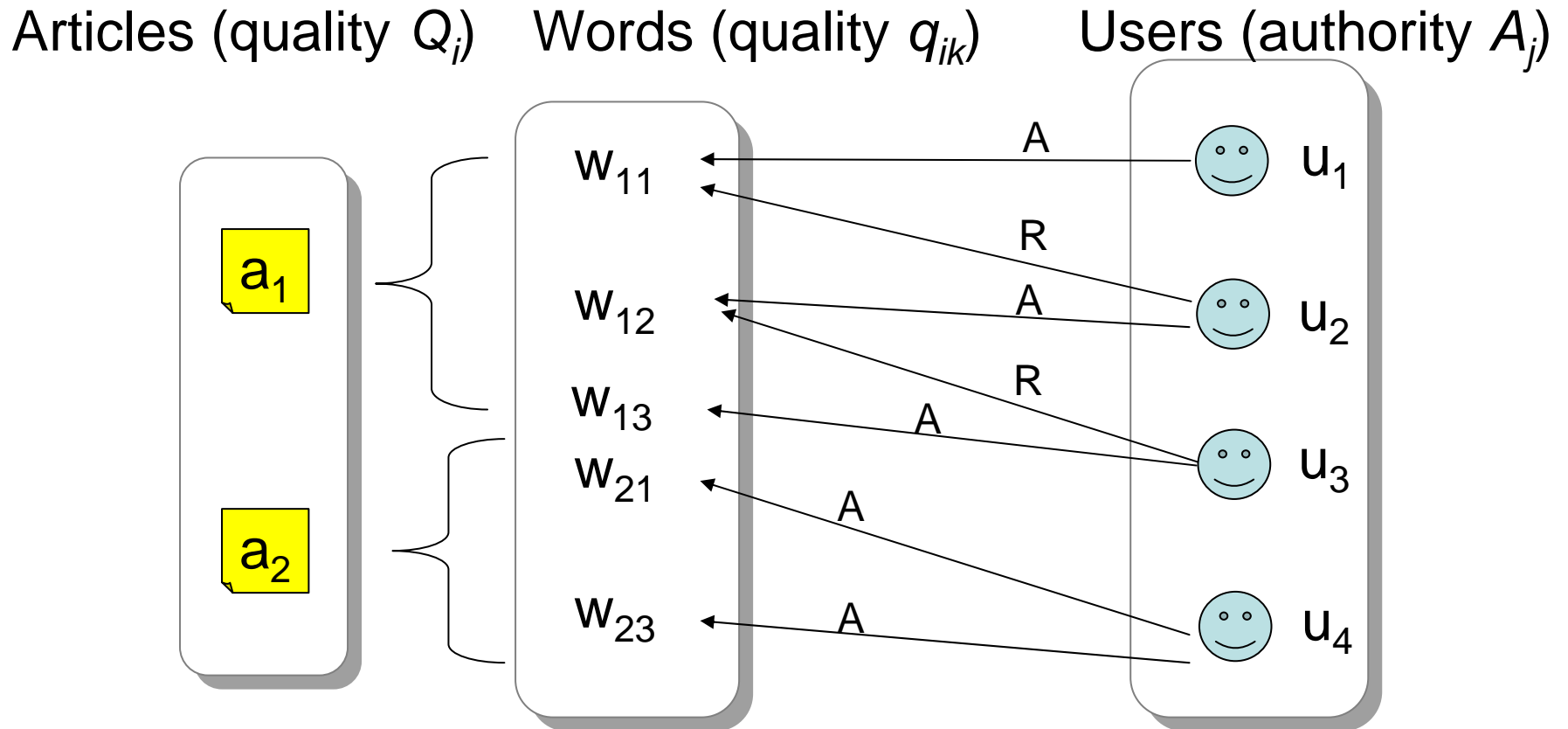
– $w_{ik} \xleftarrow{A} u_j$: w_{ik} is authored by user u_j

– $c_{ij} = \left| \left\{ w_{ik} \mid w_{ik} \xleftarrow{A} u_j \right\} \right|$

Authorship vs Reviewership

- Basic Model assumes authorship only
- Co-existence of authors and reviewers in Wikipedia
- Before an edit, a user is assumed to review the prior content of the article
- The content that survives an edit is approved by the editing user

Peer Review Model



Peer Review Model

- a_i has words $\{ w_{ik} \}$

– $w_{ik} \xleftarrow{R} u_j$: w_{ik} has been reviewed by user u_j

- Quality of word w_{ik} $q_{ik} = \sum_{\substack{(w_{ik} \xleftarrow{A} u_j) \cup \\ (w_{ik} \xleftarrow{R} u_j)}} A_j$

- Authority of user u_j $A_j = \sum_{\substack{(w_{ik} \xleftarrow{A} u_j) \cup \\ (w_{ik} \xleftarrow{R} u_j)}} q_{ik}$

- Quality of article a_i $Q_i = \sum_k q_{ik}$

Comments on Peer Review Model

- Does a user actually review the full article before editing?

May be, depending on where the user edits

- Users are:
 - More likely to review words nearby the edited text
 - Less likely to review words far away from the edited text

ProbReview Model

- Assign $w_{ik} \xleftarrow{R} u_j$ a review probability

$$\text{Prob}(w_{ik} \xleftarrow{R} u_j)$$

- Quality of word w_{ik} $q_{ik} = \sum_j f(w_{ik}, u_j) A_j$

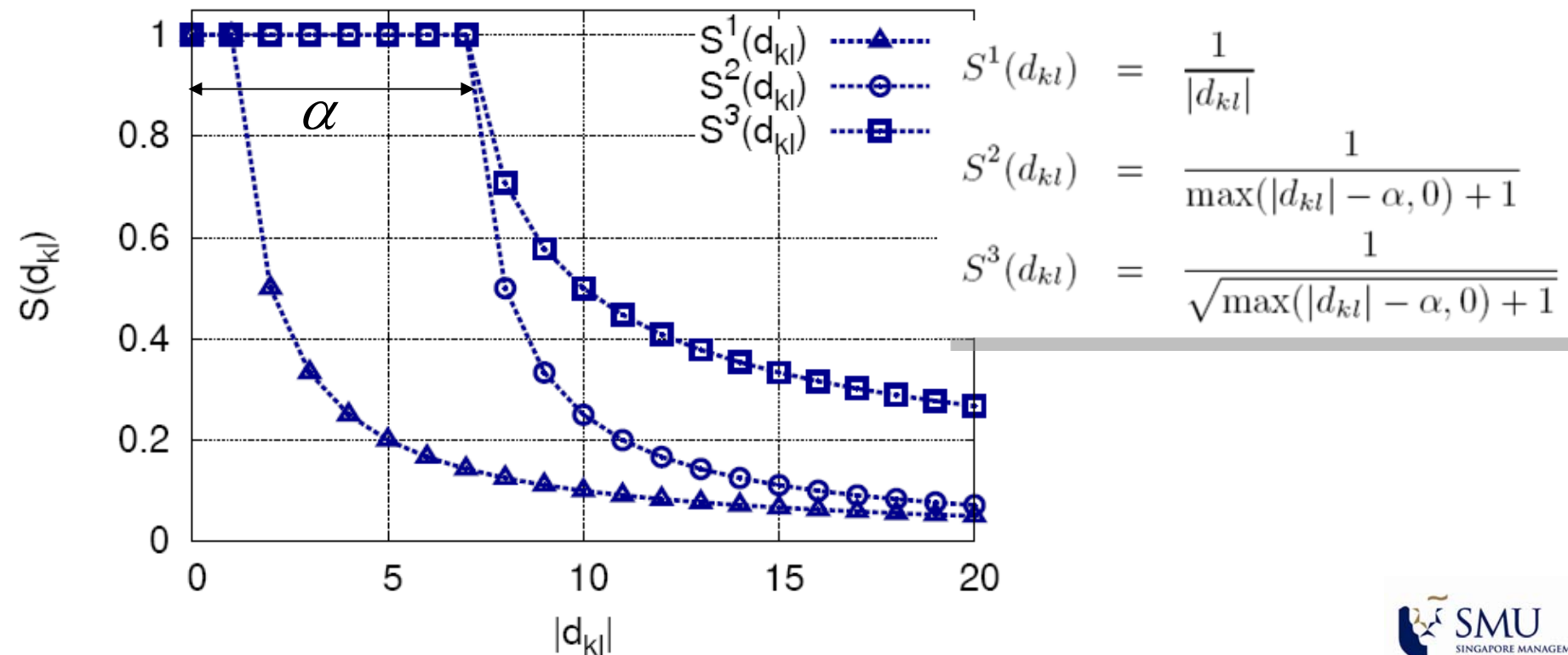
- Authority of user u_j $A_j = \sum_{i,k} f(w_{ik}, u_j) q_{ik}$

$$f(w_{ik}, u_j) = \begin{cases} 1 & \text{if } w_{ik} \xleftarrow{A} u_j \\ \text{Prob}(w_{ik} \xleftarrow{R} u_j) & \text{otherwise} \end{cases}$$

- Review probability adopts some decay scheme

Review Probability Decay Schemes

$$\text{Prob}(w_{ik} \xleftarrow{R} u_j) = \begin{cases} \max_l S(d_{kl}) & \text{if } \exists a_i^{t_m}, C(a_i^{t_m}) = u_j \wedge \\ & w_{ik} \in a_i^{t_m} \wedge \exists l, w_{il} \xleftarrow{A} u_j \\ 0 & \text{otherwise} \end{cases}$$



Computation of Basic, Peer Review, ProbReview

1. Initialize all A_j 's uniformly;
2. for each iteration:
 - Compute Q_i 's using A_j 's;
 - Compute new A_j 's using Q_i 's;
 - Normalize all A_j 's and Q_i 's by L1 norm;
3. Repeat step 2 until A_j 's and Q_i 's converges

Experiments

- Dataset: 242 country articles dated 5th Nov '06 with edit histories

Quality Class Distribution in WikiProject:Countries

Class	FA	A	GA	B	Start	Stub	Subtotal	Total
# articles	14	20	11	155	30	0	230	242
%	5.8	8.3	4.5	64.0	12.4	-	95.0	100.0
$s(p)$	4	3	2	1	0	-	-	-

Interaction Statistics from Data Processing

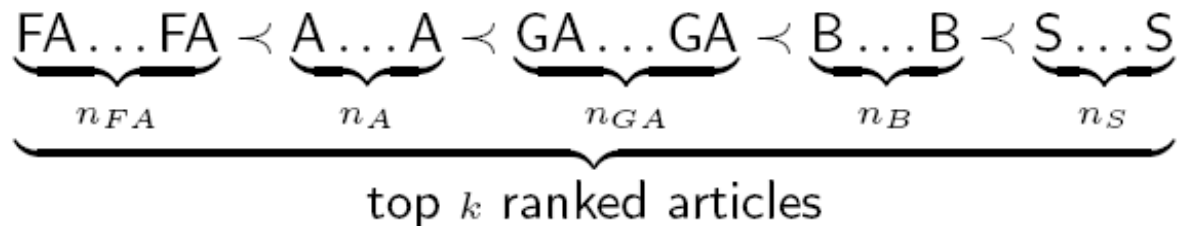
	Count	min	max	avg	std dev
# authors	per article	60	1058	227.6	138.3
# articles	per author	1	194	1.7	4.8
# words	per article	945	11,979	3,881.1	2,053.6
	per author	1	11,435	28.2	150.6
	per contribution	1	3,862	17.0	73.4
	per reviewer	0	834,572	2,437.4	10,200.2
# reviewers	per article	90	2,087	406.1	271.2
# articles	per reviewer	0	234	2.9	9.3

Performance Metric

- Normalized Discounted Cumulative Gain at top k (NDCG@ k) [0,1]

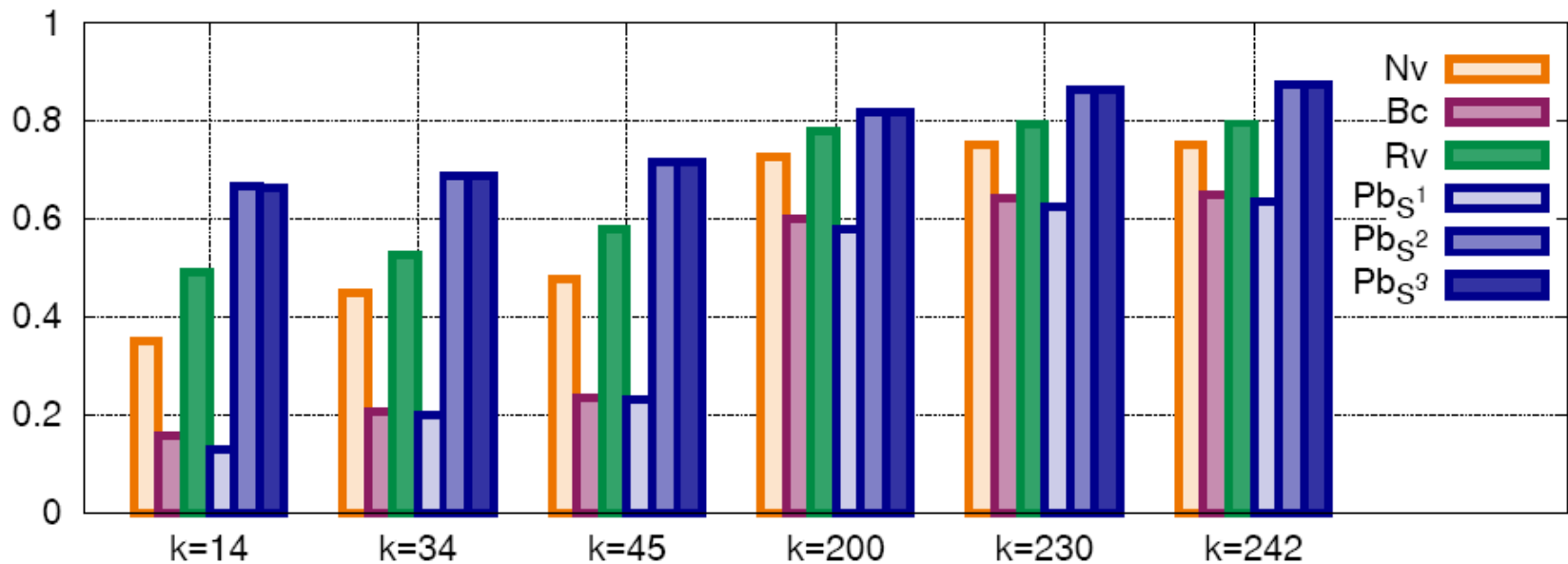
$$NDCG = \frac{1}{Z} \sum_{p=1}^k \frac{2^{s(p)} - 1}{\log(1 + p)}$$

- $s(p)$: Amount of reward to p th ranked article
- Z : Normalization factor (largest possible NDCG@ k)



NDCG@k

- ProbReview: $\alpha = 7$
(5 to 35 words in one sentence; memory span)



- Peer Review and ProbReview (s2 and s3) outperformed the others

Controversies in Wikipedia

Controversy topics in Wikipedia

- Controversy [Random House Unabridged Dictionary]
 - prolonged public dispute, debate, or contention; disputation concerning a matter of opinion.
- Similar to real life communities, controversies can arise in Wikipedia community
- Neutral Point of View (NPOV)
- Edit warring
 - “Individual editors or groups of editors repeatedly revert content edits to a page or subject area.”

Example

Disagreement in choice and rank order of sites

Request for rename/removal of article

Holiest sites in Islam - Wikipedia, the free encyclopedia - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search

Address http://en.wikipedia.org/wiki/Holiest_sites_in_Islam Go

Log in / create account

article discussion edit this page history

Holiest sites in Islam

From Wikipedia, the free encyclopedia

The examples and perspective in this article or section may not represent a **worldwide view** of the subject. Please improve this article or discuss the issue on the talk page.

There are many holy sites in the various Islamic traditions. The *Kaaba* is considered the holiest shrine, while the mosques of *Nabwi* (the Prophet) and *al-Aqsa* (farthest) are widely regarded as the second and third holiest respectively. Other shrines include the tombs of the twelve imams which are considered holy in Shiite Islam, and shrines revered by locals.

Contents [hide]

- 1 Masjid al-Haram, Mecca
- 2 Masjid-an-Nabawi, Medina
- 3 Masjid Al-Aqsa, Jerusalem
- 4 Shi'a shrines
 - 4.1 Tomb of Imam Ali, Najaf
 - 4.2 Tomb of Imam Husayn, Karbala
 - 4.3 Tomb of Imams al-Hadi and al-Askari, Samarra

Internet

Controversial topics in Wikipedia

- Controversial topics are reflection and documentation of real world
- Articles with controversial topics are likely to cause much discussions and disputes
 - Need to manually judge degree of controversy
 - Need much moderation efforts

Current solution

- Community efforts to identify controversial articles and to resolve controversy
- Manually identifying controversial articles
- Manual tagging
 - Wikipedia
 - Categories

**Behavior Mining Approach:
model article controversy
as behavior and
determine it by behavior
mining**

Tag	Meaning
{{disputed}}	The factual accuracy of this article is disputed.
{{totallydisputed}}	The neutrality and factual accuracy of this article are disputed.
{{controversial}}	This is a controversial topic, which may be under dispute.
{{disputed-section}}	Some section(s) has content whose accuracy or factual nature is in dispute.
{{totallydisputed-section}}	The neutrality and factual accuracy of some section are disputed.
{{pov}}	The neutrality of the article is disputed.

Article Tag Count (ATC)

- An estimated degree of controversy for an article

$$ATC_k = \sum_{i=1}^{N_k} c_{ki}$$

where N_k : # of revisions of article k

C_{ki} : # of controversy tags in revision i of article k

- Very few (labeled) articles have >0 ATCs
 - In our dataset, 71 out of ~20K articles have $ATC > 0$

Dataset

- 19,456 articles from the Religious Objects category and sub-categories on 12 June 2007.
- 174,338 contributors

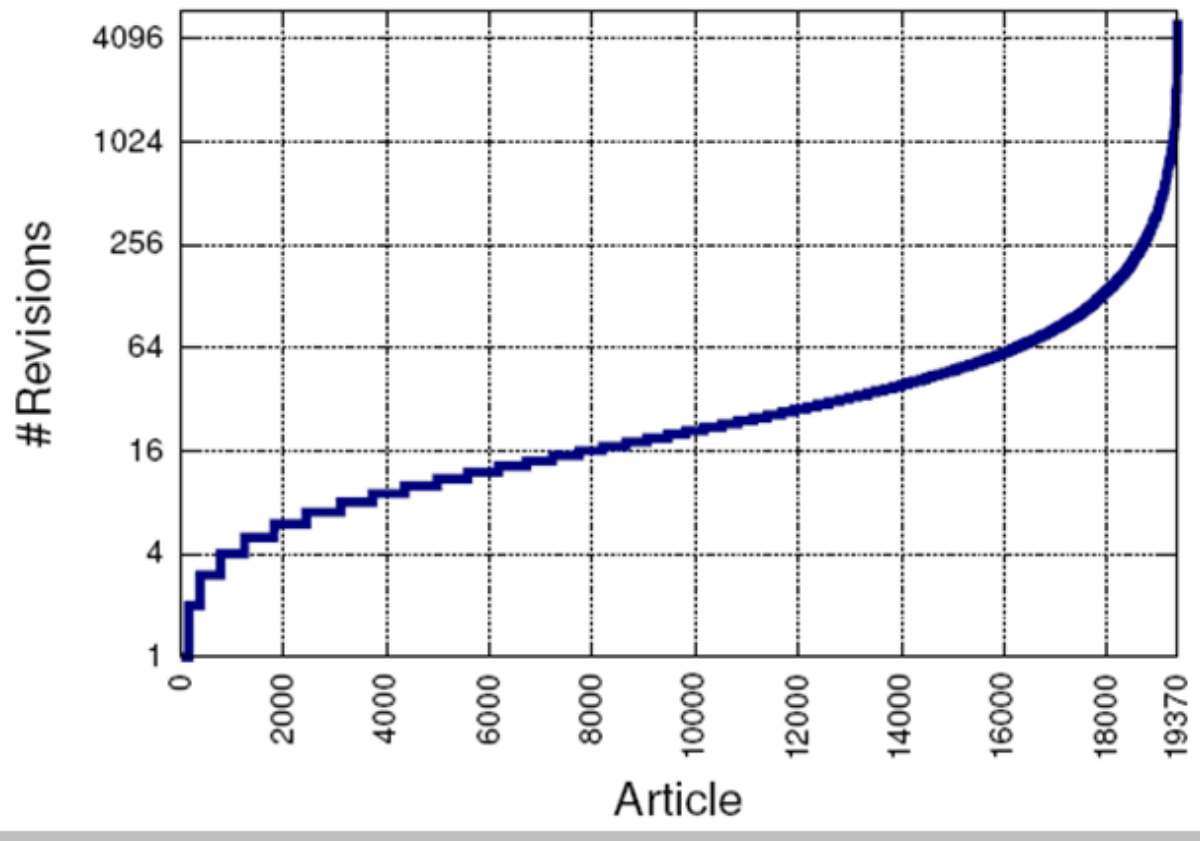
Table 4: Dataset Statistics

Count		<i>Min</i>	<i>Max</i>	Avg	Std Dev
# contributors	per article	1	3190	39.69	99.47
# articles	per contributor	1	937	2.71	23.71
Contributions	per article	3	360,929	1324.21	9118.68
	per contributor	0	347,727	140.22	3092.58
Disputes	per article	0	359,165	902.82	8863.65
	per contributor	0	348,800	95.60	3888.33

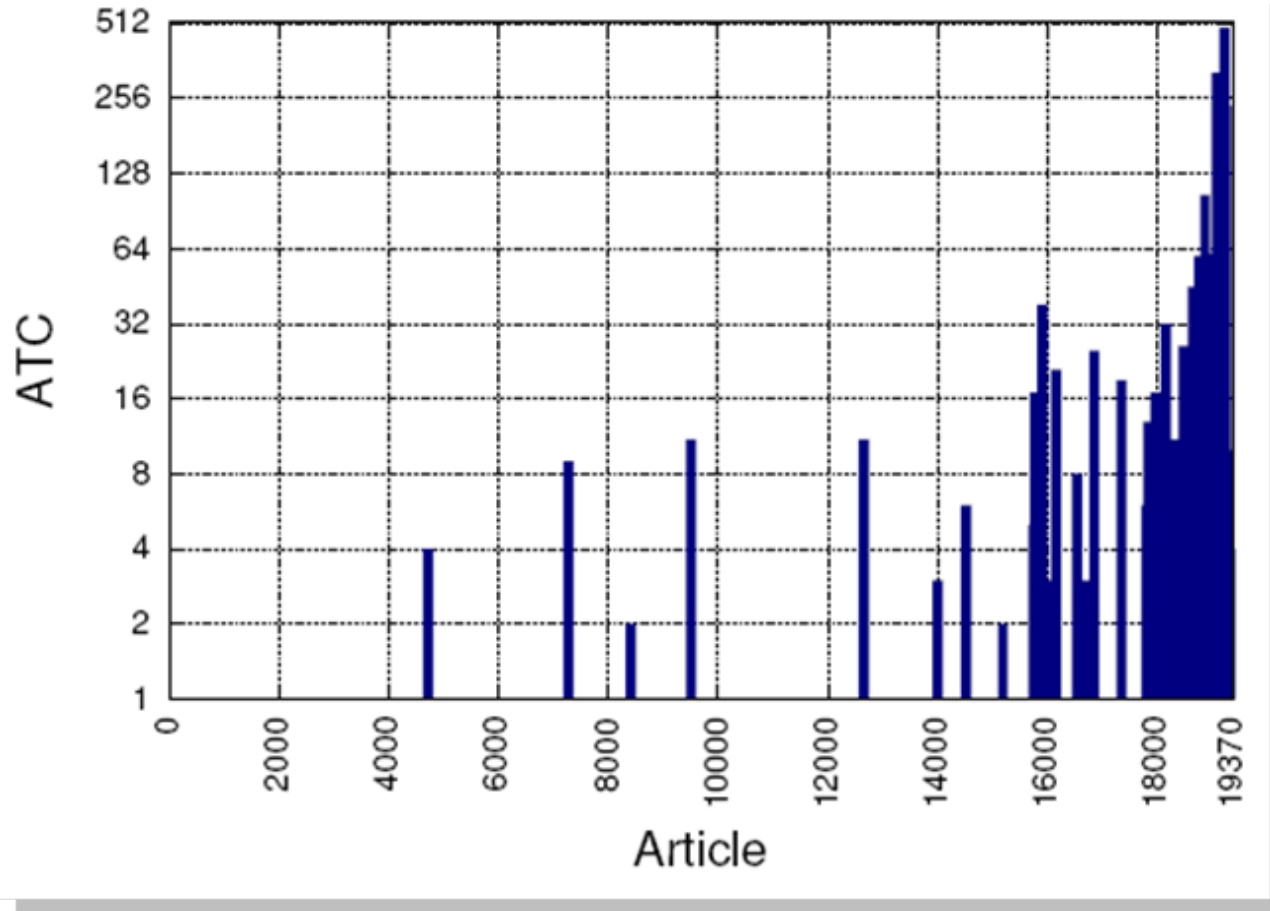
Table 5: Distribution of *ATC* Values

<i>ATC</i>	>500	101-500	21-100	5-20	1-4	0	Total
# articles	0	6	19	21	25	19,385	19,456
%	0.0	0.031	0.098	0.108	0.128	99.635	100

Articles ordered by revision count



Controversy labeled articles ordered by revision count



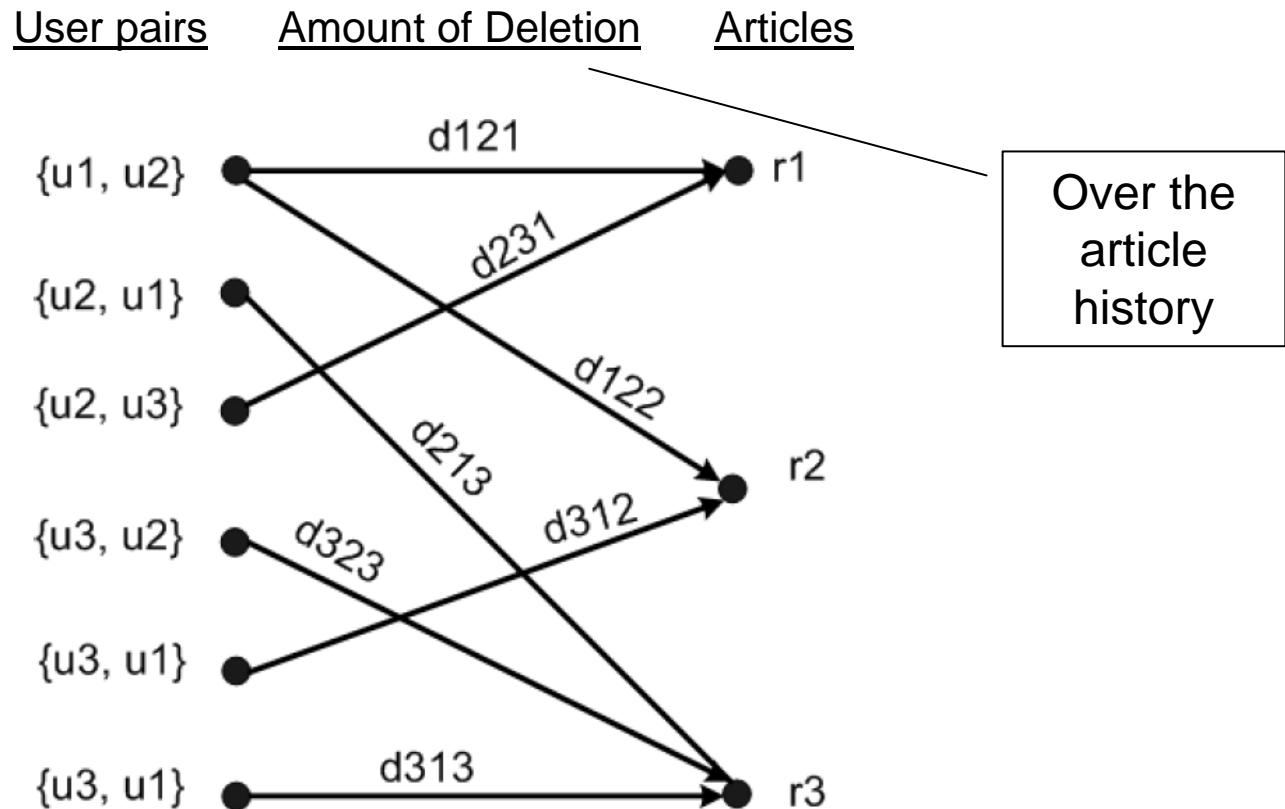
Top 20 Articles (Revision Count)

No	Article Name	#Rev	ATC	ATC Rank	No	Article Name	#Rev	ATC	ATC Rank
1	Podcast	5344	0	>71	11	Iain Lee	2384	0	>71
2	Emma Watson	4115	0	>71	12	Globe Theatre	2330	0	>71
3	Stephen Hawking	3682	0	>71	13	Grigori Rasputin	2211	0	>71
4	John McCain	3072	0	>71	14	John Howard	2200	0	>71
5	George Orwell	2948	0	>71	15	Keira Knightley	2177	0	>71
6	David Cameron	2692	0	>71	16	Salem witch trials	2176	0	>71
7	Dr.Seuss	2625	0	>71	17	Easter	2146	0	>71
8	James Madison	2623	0	>71	17	Constantine I	2111	0	>71
9	John Locke	2477	0	>71	19	Winston Churchill	2100	0	>71
10	Oscar Wilde	2432	0	>71	20	Rupert Murdoch	2003	0	>71

Top 20 Articles (Contributor Count)

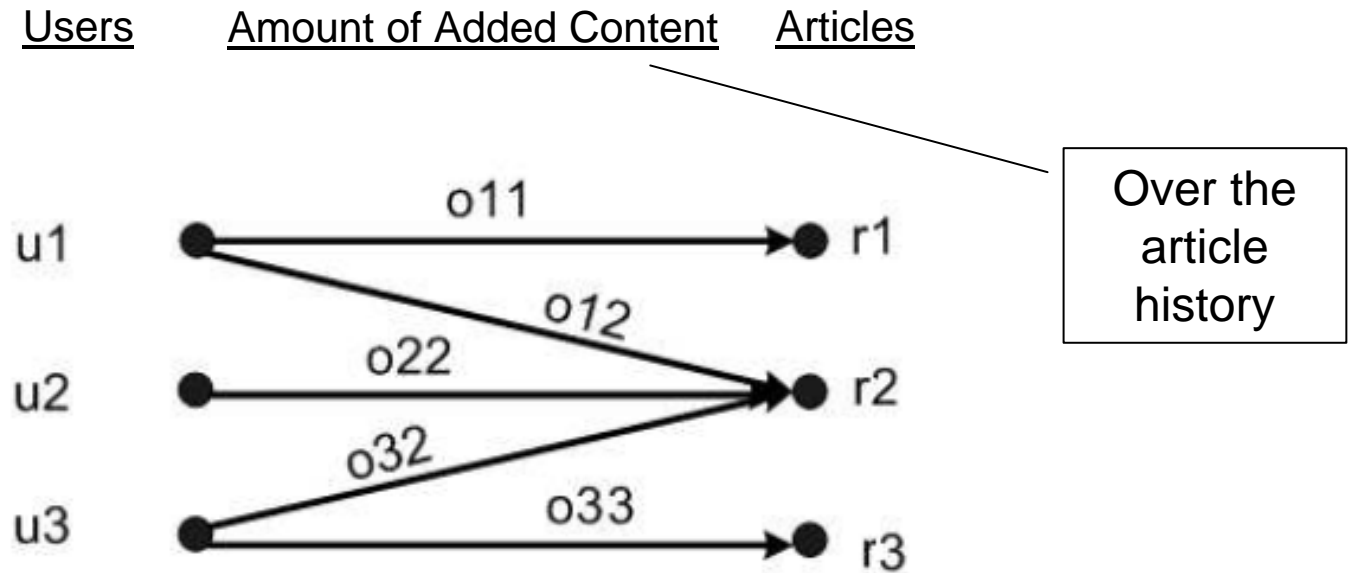
No	Article Name	#Cont	ATC	ATC Rank	No	Article Name	#Cont	ATC	ATC Rank
1	Podcast	2146	0	>71	11	Easter	1041	0	>71
2	Stephen Hawking	1933	0	>71	12	Keira Knightley	1030	0	>71
3	Emma Watson	1619	0	>71	13	Rupert Murdoch	1028	0	>71
4	John McCain	1459	0	>71	14	Globe Theatre	974	0	>71
5	George Orwell	1342	0	>71	15	Winston Churchill	948	0	>71
6	Dr. Seuss	1186	0	>71	16	David Cameron	912	0	>71
7	John Locke	1174	0	>71	17	Salem witch trials	899	0	>71
8	Grigori Rasputin	1110	0	>71	18	Jefferson Davis	879	0	>71
9	Oscar Wilde	1093	0	>71	19	Constantine I	871	0	>71
10	James Madison	1085	0	>71	20	Robert Frost	848	0	>71

Disputes in a bipartite graph



d_{ijk} = # of words contributed by user j to article k but deleted by user i

Contribution by Users



o_{ik} = # of words contributed by user i to article k

Basic model

- Controversial score of article

$$C_k^r = \frac{\sum_{i,j} d_{ijk}}{\sum_j o_{jk}}$$

- d_{ijk} = # of words contributed by user j to article k but deleted by user i
- o_{ik} = # of words contributed by user i to article k

But, is the controversy behavior of articles independent of user behaviors?

ControversyRank (CR) model

- Main idea (**mutual dependency principle**):
 - Article Controversy
An article is controversial if it contains more disputes among non-controversial contributors.
 - Contributor Controversy
A contributor is controversial if s/he is engaged in more disputes in non-controversial articles.

ControversyRank (CR) model

- Controversial score of article

$$C_k^r = \frac{\sum_{i,j} \text{agg}[(1 - C_i^u), (1 - C_j^u)] \times d_{ijk}}{\sum_j o_{jk}}$$

- Controversial score of user

$$C_i^u = \frac{\sum_{j,k} (1 - C_k^r) \times (d_{ijk} + d_{jik})}{\sum_{j,k} o_{jk} \times I(i, j, k) + \sum_k o_{ik}}$$

Choice of Agg()

- CR Average Model

$$\text{agg}[(1 - C_i^u), (1 - C_j^u)] = \frac{1 - C_i^u + 1 - C_j^u}{2}$$

- CR Product Model

$$\text{agg}[(1 - C_i^u), (1 - C_j^u)] = (1 - C_i^u) \times (1 - C_j^u)$$

Evaluation metrics

$$\text{Precision@}k = \frac{\text{\#relevant articles in top } k \text{ articles}}{k}$$

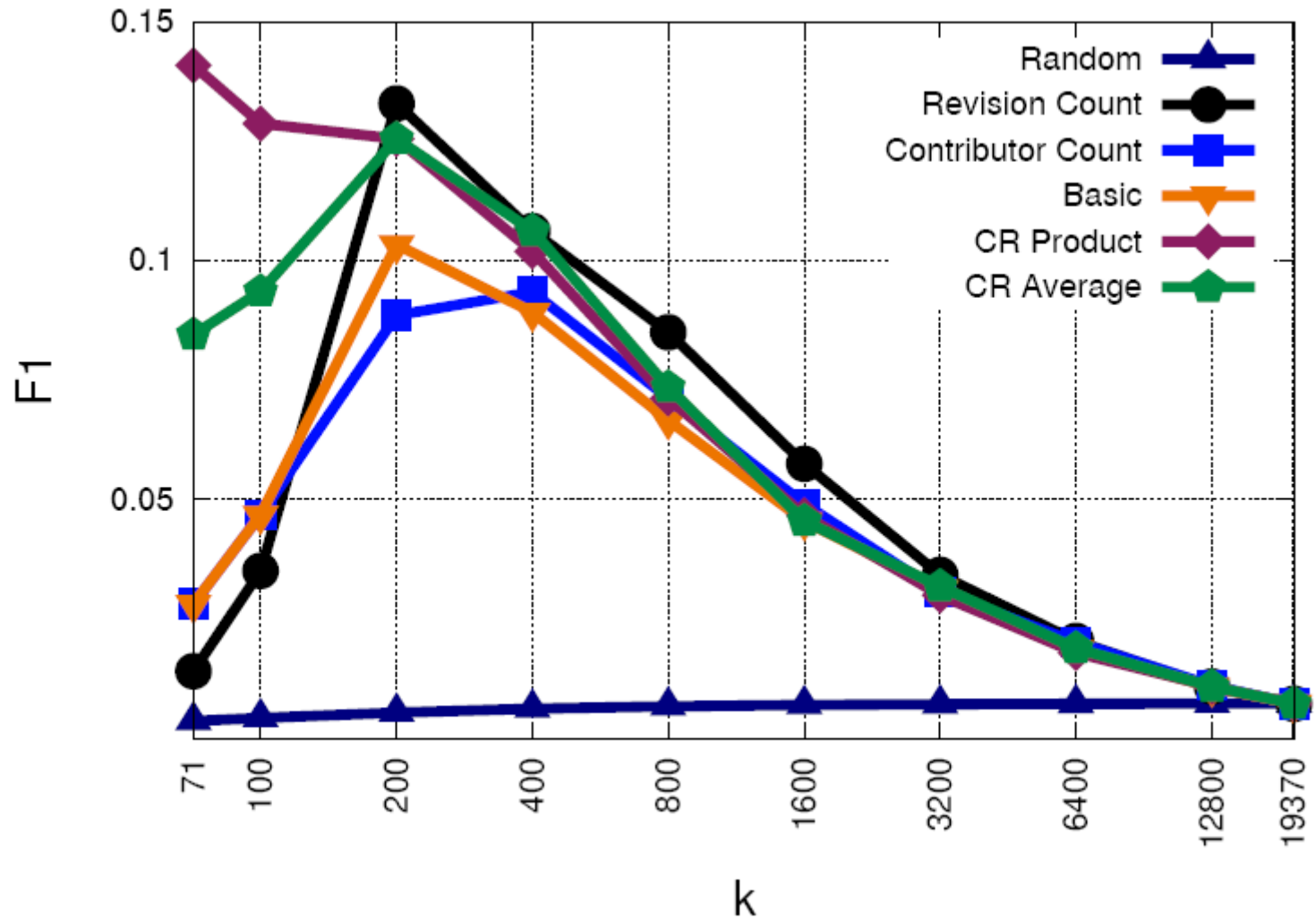
$$\text{Recall@}k = \frac{\text{\#relevant articles in top } k \text{ articles}}{\text{\#relevant articles in the dataset}}$$

$$F1@k = \frac{2 \times \text{Precision@}k \times \text{Recall@}k}{\text{Precision@}k + \text{Recall@}k}$$

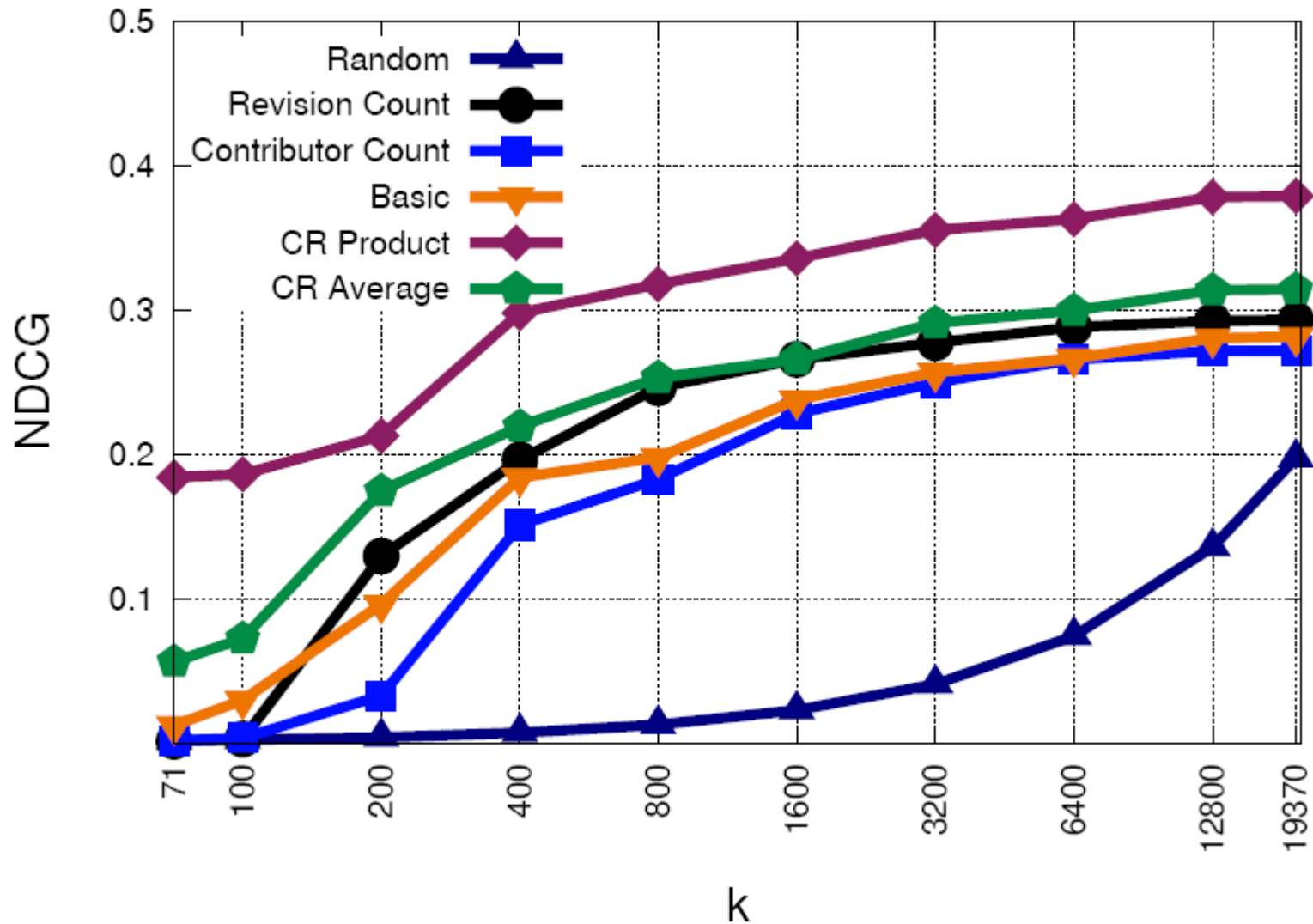
$$NDCG = \frac{1}{Z} \sum_{p=1}^k \frac{2^{s(p)} - 1}{\log(1 + p)}$$

$$s(p) = \log(ATC_p + 1)$$

Result : F1



Result : NDCG



Result: Top 20 Articles

Table 6: Basic Model Top 20 Articles

Rank	Article Name	<i>ATC</i>	<i>ATC</i> Rank	Rank	Article Name	<i>ATC</i>	<i>ATC</i> Rank
1	Dominus Illuminatio Mea	0	>71	11	John Howard	0	>71
2	North Marston	0	>71	12	Emma Watson	0	>71
3	Will Young	0	>71	13	Rozen Maiden	0	>71
4	Abingdon School	0	>71	14	Saint Sophia...	0	>71
5	Kamakhya	0	>71	15	Stephen Hawking	0	>71
6	John Dalton	0	>71	16	Queen Elizabeth II Bridge	0	>71
7	Christ Church...	0	>71	17	Aaron	0	>71
8	Podcast	0	>71	18	Kevin Rudd	0	>71
9	Jyotiba	0	>71	19	George Orwell	0	>71
10	Iain Lee	0	>71	20	Our Lady of the...	0	>71

Table 7: CR Average Top 20 Articles

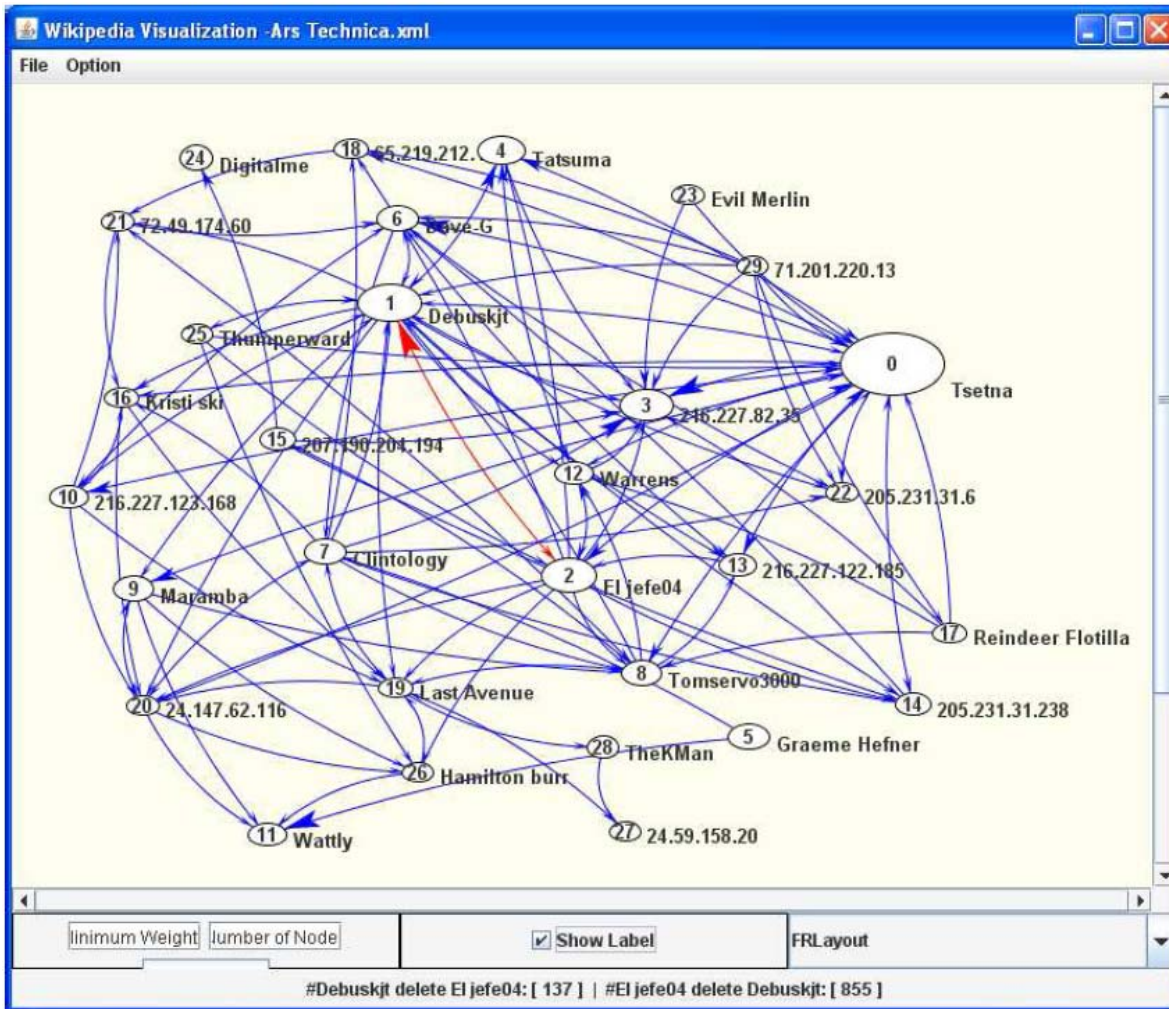
Rank	Article Name	<i>ATC</i>	<i>ATC</i> Rank	Rank	Article Name	<i>ATC</i>	<i>ATC</i> Rank
1	John Howard	0	>71	11	Globe Theatre	0	>71
2	Podcast	0	>71	12	Aaron	0	>71
3	Iain Lee	0	>71	13	Anton Chekhov	0	>71
4	Zt'l	0	>71	14	Pro-Test	58	10
5	Stephen Hawking	0	>71	15	Myron Evans	60	9
6	George Orwell	0	>71	16	Our Lady of the...	0	>71
7	Emma Watson	0	>71	17	Robert Hooke	0	>71
8	11-Sep	0	>71	18	St. Dr. Seuss	0	>71
9	Jyotiba	0	>71	19	John Dalton	0	>71
10	Andrew Adonis...	0	>71	20	John Locke	0	>71

Result: Top 20 Articles

Table 8: CR Product Top 20 Articles

Rank	Article Name	ATC	ATC Rank	Rank	Article Name	ATC	ATC Rank
1	Zt'l	0	>71	11	Bishop of Salisbury	0	>71
2	Myron Evans	60	9	12	First Baptist Church...	3	52
3	Solomon's Temple	0	>71	13	Holiest sites in Islam	490	1
4	University College Record	0	>71	14	San Lorenzo...	0	>71
5	Nassau Presbyterian Church	0	>71	15	Guy Davenport	0	>71
6	Shrine of St. Margaret...	0	>71	16	Bonn Minster	0	>71
7	Bishop of Worcester	0	>71	17	Temple Rodef Shalom	0	>71
8	Yell Group	0	>71	18	Ta Som	0	>71
9	St Volodymyr's Cathedral	0	>71	19	Romanian Orthodox...	0	>71
10	Ashtalakshmi Kovil	0	>71	20	Italo-Greek Orthodox...	0	>71

WikiNetViz: Visualizing Friends and Adversaries

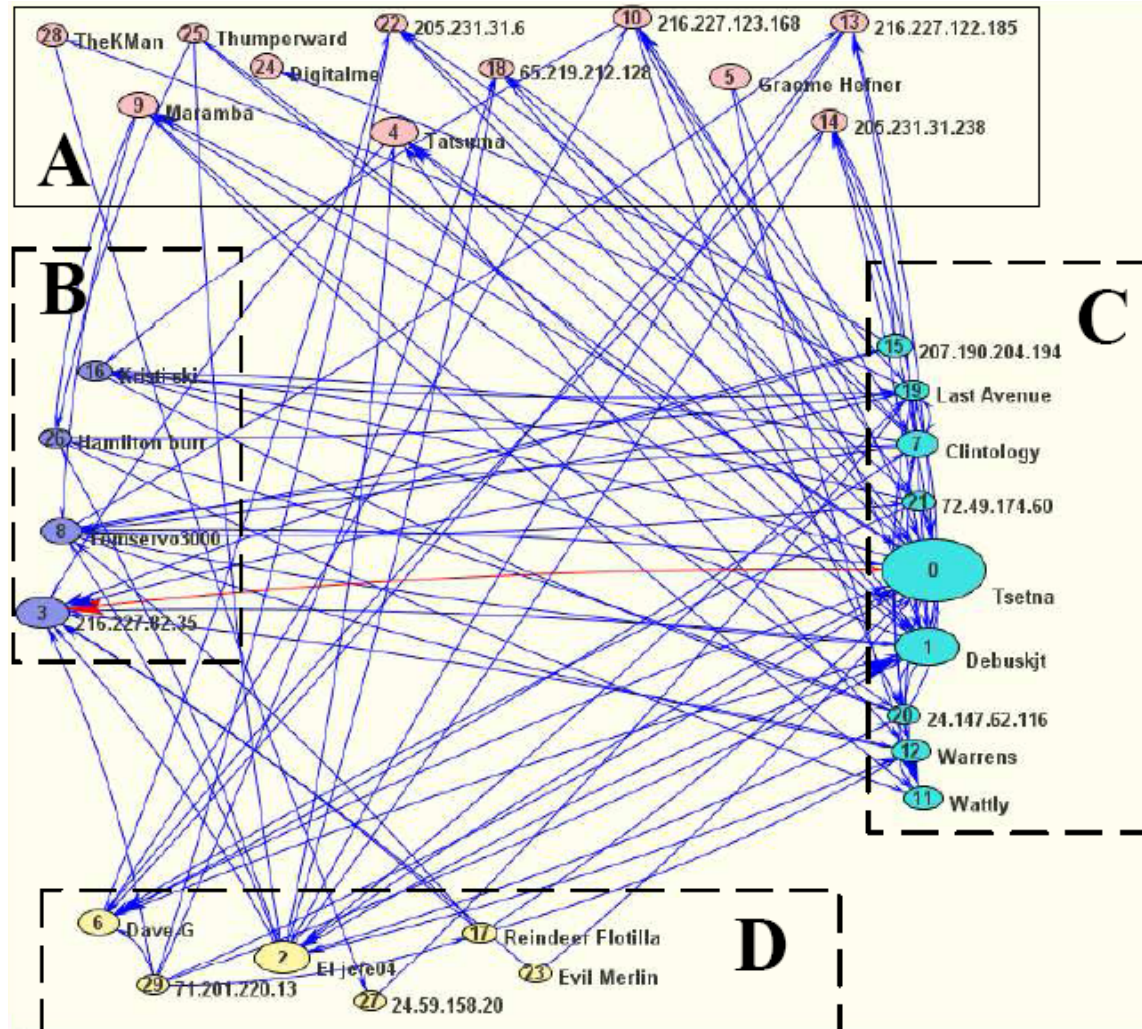


Select # contributors to be shown

Node height proportional to controversy score (CR model)

Node weight proportional to # words deleted by that user

Clustering of Users



Conclusion

Conclusion

- Interesting user/object behaviors exists in Wikipedia
- The same can be said about other kind of collaborative software systems
- Behavior mining seeks to discover such behaviors from user interactions
- Dependencies among behaviors should be considered in behavior mining
- Examples presented today include:
 - Article Quality, Article Controversy, User Authority, and User Controversy

Thank you

Ee-Peng Lim

<http://aseplim.googlepages.com/socialnetworkmining>