The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Scholarly Big Data: AI Perspectives, Challenges, and Ideas:
Technical Report WS-16-13

Automatically Augmenting Titles of Research Papers for Better Discovery

Madhavan Pallan, Biplav Srivastava
IBM Research - India

Abstract

It is well known that the title of an article impacts how well it is discovered by potential readers and read. With both people and search engines, acting on behalf of people, accessing papers from digital libraries, it is important that the paper titles should promote discovery. In this paper, we investigate the characteristics of titles of AI papers and then propose automatic ways to augment them so that they can be better indexed and discovered by users. A user study with researchers shows that they overwhelmingly prefer the augmented titles over the originals for being more helpful.

Introduction

Research papers are written to document new insights in a technical area. Once published, a paper is meant to be discovered by future researchers to build their work on and cite appropriately. Hence, one of the measures of impact of a research paper over time is its citation count that reflects both the discovery and relevance of the work to a future reader.

It is well known that the title of an article impacts how well it is discovered by potential readers and read(Kumar 2013; Falagas et al. 2013). A few previous work have noted that certain characteristics in article title correlate with their high citation count and this has lead to guidelines for authors on creating titles(Jamali and Nikzad 2011; Sagi and Yechiam 2008; Mack 2012). It is found that there exists a sweet spot for length of titles and their structure which correlate with high citiation (Habibzadeh and Yadollahie 2010). Going a step further, the authors in (Lopez, Prince, and Roche 2014) discuss algorithms to read a manuscript and come up with catchy titles with higher chance of being discovered by analyzing noun phrases in content and applying statistical criteria. Much of the work has happened here in the context of medicine(Falagas et al. 2013) although there is some interest in computer science(Anthony 2001).

Consider papers in Artificial Intelligence (AI) and Computer Science (CS) as illustrated in Table 1 from recent conferences. Some are very short (2 words) while others are extremely long (up to 19 words[1]). The short ones may be easy to remember but too narrow for discovery over time. The long ones may also be too verbose for search engines and humans to work with. One paper in particular, #12 from WWW-12, is catchy but does not convey much about the scope of the paper, its methods or contributions - all desirable for paper titles. To address these gaps, we are interested in effective and efficient ways of arriving at title of a paper so that it can be easily discovered manually or automatically from a digital library. Our considerations are that (a) the authors should continue to retain the actual control on titles just as they do for the rest of the paper's content, and (b) the title information, if needed, can be automatically augmented based on public metadata so that they can be easily discovered. The metadata can come from how the paper is presented (e.g., session names) or published (e.g., ACM's classification system). However, metadata that goes into paper submission systems (like keywords and tracks) but not retained during presentation or publishing are unavailable.

In the remainder, we begin by reviewing guidelines for what makes for good paper titles. Then we motivate the paper by looking at the statistics of AI paper titles at major conferences and establishing the need to improve them, and finding a relationship between titles of CS/ AI papers and their citations. We then explore different ways the titles can be augmented based on where the paper is presented and published in a digital library. We conduct a user study to validate the usefulness of augmented titles and finally conclude by providing a prescription for how papers titles should be published to improved manual and automatic discovery.

The contributions in the paper are that we:

1. assess the characteristics of paper titles in AI and CS conferences and establish need to improve them for discovery
2. evaluate relationship between title lengths of CS papers with citations for well-established papers
3. propose augmentation of titles with additional metadata and experiment with two different augmentation approaches - session names and ACM CCS, and
4. prescribe how to augment paper titles for discovery without burdening authors

It is noteworthy that although there are previous works which have studied syntactic features of paper titles like word count, punctuation mark and structure, this is the first

[1]As explained later, we consider titles of papers as they appear in a conference's schedule.

S.No.	Conference	Length	Paper Title
1	AAAI-12	2	Solving Dots-And-Boxes
2	AAAI-12	2	Weighted Clustering
3	AAAI-12	15	Parsing Outdoor Scenes from Streamed 3D Laser Data Using Online Clustering and Incremental Belief Updates
4	AAAI-12	18	Design and Optimization of an Omnidirectional Humanoid Walk: A Winning Approach at the RoboCup 2011 3D Simulation Competition
5	AAAI-13	2	Clustering Crowds
6	AAAI-13	2	Uncorrelated Lasso
7	AAAI-13	17	Online Optimization with Dynamic Temporal Uncertainty : Incorporating Short Term Predictions for Renewable Integration in Intelligent Energy Systems
8	AAAI-13	19	Computational Sustainability Award Winner: A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings
9	WWW-12	4	Distributed Graph Pattern Matching
10	WWW-12	4	Crowdsourcing with Endogenous Entry
11	WWW-12	16	Counting Beyond a Yottabyte, or how SPARQL 1.1 Property Paths will Prevent Adoption of the Standard
12	WWW-12	17	It's Simply Integral to What I do: Enquiries into how the Web is Weaved into Everyday Life
13	WWW-13	2	Reactive Crowdsourcing
14	WWW-13	3	Content-Aware Click Modeling
15	WWW-13	17	Is This App Safe for Children? A Comparison Study on Maturity Ratings on Android and iOS Applications
16	WWW-13	18	I Know the Shortened URLs You Clicked on Twitter: Inference Attack using Public Click Analytics and Twitter Metadata

Table 1: Illustration of shortest and longest paper titles from recent AI and CS conferences.

work for AI and CS which also proposes ways to improve titles and evaluates them in a user study to arrive at a prescription.

What Makes For a Good Paper Title?

We start by reviewing guidelines for what makes for good paper titles. The literature gives a diverse spectrum of views. The authors in (Lopez, Prince, and Roche 2014) note that a title is syntactically a *metadata with a structure that can be a word, phrase, expression, or sentence, that serves to indicate a paper or one of its parts and give its subject*. In essence, it could be any string but the expectation is for it to be meaningful with respect to its content. In terms of what a title should contain, (Kumar 2013) argues that the title should be captivating and introduce the reader to the subject of the paper in a clear and concise manner. But what is concise? A prominent university's guideline (USC 2014) is to have 10-15 words while a survey of computer science papers (Anthony 2001) finds them to be from anywhere in 2-24 range with an average of 9. There are 3 main types of paper titles: nominal, which describes a paper's theme; compound, which has title made up of multiple parts separated by a punctuation mark, with each part standing for a specific aspect of the paper; and full sentences that assert a major contribution by the paper(Kumar 2013). Nominal and compound titles are common in computer science.

We note that a paper's title are used by two different user groups: (a) humans who search for prior work in digital and print media, and (b) automated routines in modern digital libraries which use them to build indices to support effective search interfaces. Humans users prefer short, catchy, precise titles which they can remember and recall. However, since words have many synonyms and technical areas evolve over time, short terms can hamper discovery and degrade search. For automated search, it is known in asset organization literature that descriptive titles which contain the context of usage are preferable(Srivastava et al. 2007). But this invariably results in longer asset metadata, i.e., titles here.

For the purpose of the paper, we adopt the guidelines for paper titles to be that they:

- should be between 5-15 words long
- should have standard terms that help automatic indexing
- should convey the subject of the paper

S.No.	AAAI -2012	AAAI -2013	WWW -2012	WWW -2013
Papers	294	254	107	125
Titles	(2,18 ,7.918)	(2,19 ,8.244)	(4,17 ,8.355)	(2,18 ,9.104)
Titles (w/o SW)	(2,13) ,6.316	(2,15) ,6.5	(2,13) ,6.551	(1,15) ,7.096
Unique Words	1090	993	512	586
Unique Words (w/o SW)	1074	976	496	568

Table 2: Overall statistics about papers in the dataset. Numbers in bracket represent (min,max,avg) and w/o SW denotes without stop words.

- may be catchy to help humans recall it later

Paper Titles in CS and AI

We now investigate the nature of paper titles in major conferences and their characteristics in highly cited papers. This will set the stage for motivating automatic augmentation of paper titles for discovery.

Paper Titles in CS/ AI Conferences

We considered the papers published in two major conferences for two years (2012-2013): Association for the Advancement for Artificial Intelligence (AAAI) conference, which is the premier AI conference(AAAI 2012; 2013), and World Wide Web (WWW) conference(WWW 2012; 2013), where papers appear from both AI and other CS disciplines. The information about papers in these conferences are available at multiple fora but with different, overlapping, metadata. Since we were interested in both paper details and how they were grouped together for presentation, we used the cited online sources. Here, the titles of the papers are extracted and then analyzed as-is and also after extracting common words called stopwords[2]. The tools used were JSoup[3] for extraction and R for statistical analysis[4].

Table 2 gives the overall statistics for the papers appearing in these conferences. The total number of papers are 780 which is more than papers analyzed in previous studies like (Anthony 2001) (600 papers). The title lengths varied from 2-19 with an average of about 8 words, not accounting for stop words, and 1-15 with an average of about 7 words when they are considered. The most frequent 5 words in the titles are shown in Table 3. They reflect the AI and web focus of the AAAI and WWW conferences, respectively.

Paper Titles and Citations

Since papers are written to be discovered and cited, we now look at the nature of some of the highest cited papers in computer science and AI. To do so, we obtained the authors and citation counts from Google Scholar[5] as on 13 August 2014. In Figure 1, title lengths of paper by the 10 highest cited authors in CS are shown along with the total. The plots have titles with and without commonly used words (stop words), in left and right, respectively.

In Figure 3 and Table 4, the citation data for the top-10 authors and total are fitted to a sinusoidal function - $a_1 sin(b_1 x + c_1) + a_2 sin(b_2 x + c_2)$. Also shown is the R-Square values which is a statistical measure showing how successful the fit is in explaining the variation of the data. Mathematically, it is defined as the ratio of the sum of squares of the regression (SSR) and the total sum of squares (SST). R-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. Thus, an R-square value of 0.8907 means that the fit explains 89.07% of the total variation in the data about the average. The R-squared value shows that the curve fit to sinusoidal is, in general, quite good. Further, the highly cited papers were typically of length 2-10 words.

Discussion

From the papers in two major conferences over 2 years, we found that the papers can have a wide diversity in title length (2-19) and an average of about 8 words. To the extent highly cited papers provide clues to how to name papers, they suggest short titles of length 2-10 words. Finally, guidelines on writing good paper titles recommend using 5-15 words which are relevant to the scope of the paper.

We now consider how to augment paper titles. This would be especially helpful for the case where the title is too short or not relevant to its material, but we still want to promote discovery.

Automatically Augmenting Paper Titles

In this section, we show how titles can be automatically augemented to promote meaningful discovery. We refer to the orginal paper titles as *Original*. The new content comes from the conferences' session names and ACM classification of Computer Science discipline, which are sometimes different but metadata authors put in conference paper submission systems but are not retained thereafter.

Augmenting Titles with Session Names

After research papers are accepted for publication, they are scheduled for presentation. Conferences differ in how they group papers into different sessions for oral or poster presentation or both, but are uniform in assigning additional metadata to label the groups in which the papers get presented. We use the term *session name* to refer to presentation group names. Since program organizers have already considered the paper's content and reviews to group them and assign meaningful names, they are good candidates for metadata to augment paper titles.

For our work, we used the schedule of technical program of conferences as our data source. The Jsoup library came

[2]Stopwords used are: A, AN, THE, FOR, AND, NOR, BUT, OR, YET, SO, IN, ON, OF, UNDER, ABOVE, ABOUT, AT, BEHIND, FROM, OPPOSITE, UPON, TILL, THROUGH, SINCE, AROUND, ALONG, INTO, WITH, BY, TO.

[3]http://www.jsoup.org

[4]http://www.r-project.org/

[5]http://www.googlescholar.com

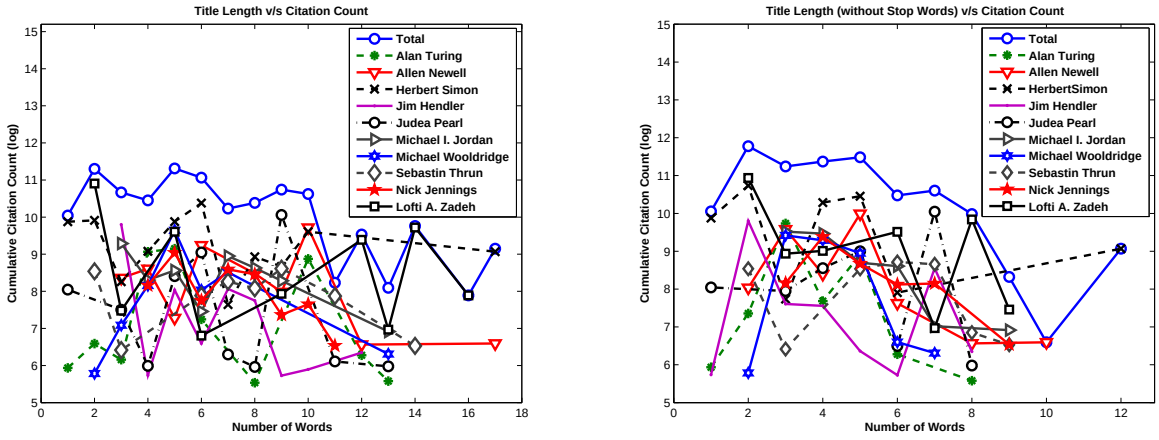

Figure 1: Length of paper titles and their citations

S.No.	AAAI-2012	AAAI-2013	WWW-2012	WWW-2013
1	LEARNING (33, 0.112)	LEARNING (34, 0.134)	WEB (18, 0.168)	SOCIAL (21, 0.168)
2	USING (25, 0.085)	PLANNING (14, 0.055)	SOCIAL (15, 0.14)	WEB (19, 0.152)
3	GAMES (17, 0.058)	SEARCH (14, 0.055)	SEARCH (13, 0.121)	NETWORKS (15, 0.12)
4	MODELS (15, 0.051)	MODEL (11, 0.043)	NETWORKS (10, 0.093)	SEARCH (12, 0.096)
5	APPROACH (14, 0.048)	VIA (11, 0.043)	INFORMATION (7, 0.065)	ONLINE (11, 0.088)

Table 3: Top 5 words in the Respective Conference. Numbers in bracket represent (#papers, fraction of papers) statistics from conferences.

very handy in extracting the information. The extracted session names were used as a prefix to the original titles marked in square brackets. We refer to such paper titles as *Aug-Session*.

Augmenting with ACM Classification

Association for Computing Machinery, ACM, the leading computing organization, had developed a classification scheme for computer science called Computing Classification System (CCS) in 1964 and has since been revising and maintaining it frequently. The latest version is called 2012 CCS (ACM 2012) and it is a poly-hierarchical ontology used extensively in ACM's own digital library. 2012 CCS can be used for educational and research purposes. Of the two conferences considered, WWW is an ACM conference while AAAI is not.

In order to augment papers titles with ACM 2012 CCS terms, the steps followed were:

1. **Match:** Take session name of papers and match it to 2012 CCS automatically.
2. **Select:** Select metadata from ACM node and children
3. **Build:** Decide and build the augmented name

To match session names with ACM CCS, we resort to string comparison techniques. Let $\theta(s_i, s_j)$ represent the similarity score of two single contents (strings), s_i and s_j. There are many choices for θ (Cohen, Ravikumar, and Fienberg 2003; Chapman 2009). We experimented with Monge-Elkan, Levenshtein, Cosine and Euclidean. Now s_i is similar to s_j, denoted by $s_i \approx s_j$, if $\theta(s_i, s_j) \succ \phi$, where ϕ is some threshold. In our experiments, we found Levenshtein and Cosine to perform the best. Details are omitted due to page restrictions but we point out a complication. It is how to navigate the CCS poly-hierarchy because a match to a node may also match its specialized children. We navigated the CCS breadth (level-) first and stopped at the first node that matched.

Once a matching ACM CCS node is found, one could select its label as well as one of more of its descendents. In order to keep the augmented titles brief, we chose to select the names of only the matching node and leave exploration of other choices as future work.

After selecting the ACM metadata, we build the augmented name by prefixing it to the original name. The prefix is marked in square brackets. We refer the paper titles augmented with ACM 2012 CCS as *Aug-ACM*.

Discussion

Figure 2 shows the characteristics of the paper titles for the full dataset for *Original*, *Aug-Session* and *Aug-ACM* with a lognormal probability density function of the respective # words in the title. As expected, we see the average title lengths increasing compared to original with *Aug-ACM* leading to a lesser increase than *Aug-Session*. Since ACM is manually curated as an ontology, the augmented metadata is expected to be more meaningful as well.

Further, we show the automatic augmentation of the example titles from Table 1 in Table 5. We wanted to see whether researchers actually prefer augmented paper titles over original and if so, which among the two we investigated. To do so, we conducted a user study that is described

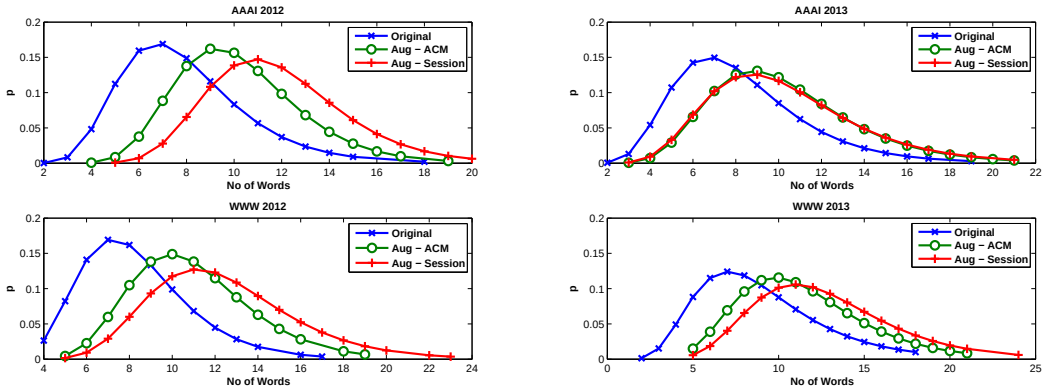

Figure 2: Augementing titles of papers with session names and ACM classification; latter is shorter.

S.No.	Person Name	Total Citations	R-Squared Value	RMSE	RMSE % Error
1	Allen Newell	47,261	0.8907	2806	5.9%
2	Alan Turing	29,217	0.8695	2076	7.1%
3	Nicks Jennings	28,713	0.8622	1743	6.1%
4	Sebastin Thrun	24,287	0.6889	1884	7.7%
5	Jim Hendler	28,929	0.682	5242	18.1%
6	Michael Wooldridge	32,444	0.6696	5237	16.1%
7	Herbert Simon	142,985	0.6498	7608	5.3%
8	Michael I.Jordan	40,101	0.583	3823	9.5%
9	Lofti A. Zadeh	107,185	0.5814	97	∼0%
10	Judea Pearl	43,398	0.541	7277	16.7%
	Total	524,520	0.6023	107	∼0%

Table 4: Top 10 authors in CS, their papers' citation counts and the error values for fit obtained for their papers.

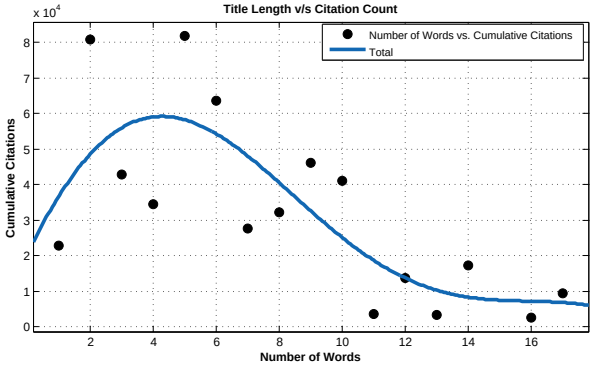

Figure 3: Sinusoidal curve fitted to plot relating total citations to title lengths.

next.

User Study and Discussion

We conducted an online survey with a random group of participants consisting of researchers and advanced developers of an industrial research organization working in Information Technology-based solutions and services. The AI research experience of participants ranged from 1-20 years. A total of 15 participants participated in the survey. The survey gave a single instruction, *"In the following list of paper titles, click on the style of paper title given, which you consider the most meaningful if you were to search for them"*. It consisted of paper title set consisting of 16 titles in which we took 4 titles each from conferences AAAI- 2012, AAAI-2013, WWW-2012 and WWW-2013.There were two types of titles. The first type (Type-1) consisted of 8 titles, each having three options to select from, which were research title (*Original*), research title augmented with session track (*Aug-Session*) and research title augmented with ACM 2012 CCS (*Aug-ACM*). The second type (Type-2) had 8 titles, each having only 2 options since *Aug-Session* turned out to be the same as *Aug-ACM* . Participants were not told about the type of titles to avoid any bias in the reply. For each type, the total number of responses would be 120 (i.e., = 8 x 15).

The result is shown in Figure 4. Here, the count of votes of respondents is shown on the Y-axis for Type-1 and Type-2 titles on left and right, respectively. In Type-1, we see that 75% of the responses preferred augmented titles over the original (25%), and this was made up of 42% preferring *Aug-ACM* and 33% liking *Aug-Session*. In Type-2, we see that 62% of the responses preferred augmented titles over the original (38%).

The user study, although limited in size, shows that augmented paper titles are preferrable than original titles for discovery, and among the two types of augmentation, adding

S.No.	Conference	Paper Title	Paper Title Augmented with Session Name	Paper Title Augmented with ACM CCS
1	AAAI-2012	Solving Dots-And-Boxes	[Multiagent Systems] Solving Dots-And-Boxes	[Multiagent Systems] Solving Dots-And-Boxes
2	AAAI-2012	Weighted Clustering	[Machine Learning] Weighted Clustering	[Machine Learning] Weighted Clustering
3	AAAI-2012	On Maxsum Fair Cake Divisions	[Multiagent Systems] On Maxsum Fair Cake Divisions	[Multiagent Systems] On Maxsum Fair Cake Divisions
4	AAAI-2012	Don't Be Strict in Local Search!	[Constraints, SAT, & Search] Don't Be Strict in Local Search!	[Search interfaces] Don't Be Strict in Local Search!
5	AAAI-2012	Parsing Outdoor Scenes from Streamed 3D Laser Data Using Online Clustering and Incremental Belief Updates	[Robotics] Parsing Outdoor Scenes from Streamed 3D Laser Data Using Online Clustering and Incremental Belief Updates	[Robotics] Parsing Outdoor Scenes from Streamed 3D Laser Data Using Online Clustering and Incremental Belief Updates
6	AAAI-2012	Design and Optimization of an Omnidirectional Humanoid Walk: A Winning Approach at the RoboCup 2011 3D Simulation Competition	[Robotics] Design and Optimization of an Omnidirectional Humanoid Walk: A Winning Approach at the RoboCup 2011 3D Simulation Competition	[Robotics] Design and Optimization of an Omnidirectional Humanoid Walk: A Winning Approach at the RoboCup 2011 3D Simulation Competition
7	AAAI-2013	Uncorrelated Lasso	[Classification] Uncorrelated Lasso	[Classification and regression trees] Uncorrelated Lasso
8	AAAI-2013	Clustering Crowds	[Crowdsourcing] Clustering Crowds	[Crowdsourcing] Clustering Crowds
9	AAAI-2013	Search More, Disclose Less	[Privacy and Social Media] Search More, Disclose Less	[Social media] Search More, Disclose Less
10	AAAI-2013	Teaching Classification Boundaries to Humans	[Classification] Teaching Classification Boundaries to Humans	[Classification and regression trees] Teaching Classification Boundaries to Humans
11	AAAI-2013	Online Optimization with Dynamic Temporal Uncertainty : Incorporating Short Term Predictions for Renewable Integration in Intelligent Energy Systems	[MDPs and Sequential Processes] Online Optimization with Dynamic Temporal Uncertainty : Incorporating Short Term Predictions for Renewable Integration in Intelligent Energy Systems	[Markov decision processes] Online Optimization with DynamicTemporal Uncertainty : Incorporating Short Term Predictions for Renewable Integration in Intelligent Energy Systems
12	AAAI-2013	Computational Sustainability Award Winner: A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings	[Data Mining] Computational Sustainability Award Winner: A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings	[Data Mining] Computational Sustainability Award Winner: A Temporal Motif Mining Approach to Unsupervised Energy Disaggregation: Applications to Residential and Commercial Buildings
13	WWW-2012	Crowdsourcing with Endogenous Entry	[Crowdsourcing] Crowdsourcing with Endogenous Entry	[Crowdsourcing] Crowdsourcing with Endogenous Entry
14	WWW-2012	Distributed Graph Pattern Matching	[Data And Content Management] Distributed Graph Pattern Matching	[Content analysis and feature selection] Distributed Graph Pattern Matching
15	WWW-2012	Trains of Thought: Generating Information Maps	[Web Mining] Trains of Thought: Generating Information Maps	[Web Mining] Trains of Thought: Generating Information Maps
16	WWW-2012	Your Two Weeks of Fame and your Grandmother's	[Web Mining] Your Two Weeks of Fame and your Grandmother's	[Web Mining] Your Two Weeks of Fame and your Grandmother's
17	WWW-2012	Counting Beyond a Yottabyte, or how SPARQL 1.1 Property Paths will Prevent Adoption of the Standard	[Ontology Representation and Querying : RDF and SPARQL] Counting Beyond a Yottabyte, or how SPARQL 1.1 Property Paths will Prevent Adoption of the Standard	[Ontology engineering] Counting Beyond a Yottabyte, or how SPARQL 1.1 Property Paths will Prevent Adoption of the Standard
18	WWW-2012	It's Simply Integral to What I do": Enquiries into how the Web is Weaved into Everyday Life"	[User Interfaces and Human Factors] It's Simply Integral to What I do": Enquiries into how the Web is Weaved into Everyday Life"	[User studies] It's Simply Integral to What I do": Enquiries into how the Web is Weaved into Everyday Life"
19	WWW-2013	Reactive Crowdsourcing	[Social Web Engineering] Reactive Crowdsourcing	[Web and social media search] Reactive Crowdsourcing
20	WWW-2013	Content-Aware Click Modeling	[Web Mining] Content-Aware Click Modeling	[Web Mining] Content-Aware Click Modeling
21	WWW-2013	Pick-A-Crowd: Tell Me What You Like, and Ill Tell You What to Do	[Social Web Engineering] Pick-A-Crowd: Tell Me What You Like, and Ill Tell You What to Do	[Web and social media search] Pick-A-Crowd: Tell Me What You Like, and Ill Tell You What to Do
22	WWW-2013	Is It Time for a Career Switch	[Recommender System] Is It Time for a Career Switch	[Recommender Systems] Is It Time for a Career Switch
23	WWW-2013	Is This App Safe for Children? A Comparison Study on Maturity Ratings on Android and iOS Applications	[Transforming UIs / Personal & Mature Data] Is This App Safe for Children? A Comparison Study on Maturity Ratings on Android and iOS Applications	[User interface programming] Is This App Safe for Children? A Comparison Study on Maturity Ratings on Android and iOS Applications
24	WWW-2013	I Know the Shortened URLs You Clicked on Twitter: Inference Attack using Public Click Analytics and Twitter Metadata	[Privacy and Personalization] I Know the Shortened URLs You Clicked on Twitter: Inference Attack using Public Click Analytics and Twitter Metadata	[Personalization] I Know the Shortened URLs You Clicked on Twitter: Inference Attack using Public Click Analytics and Twitter Metadata

Table 5: Select papers with original titles as well as when augmented with conference's session names and ACM classification, respectively.

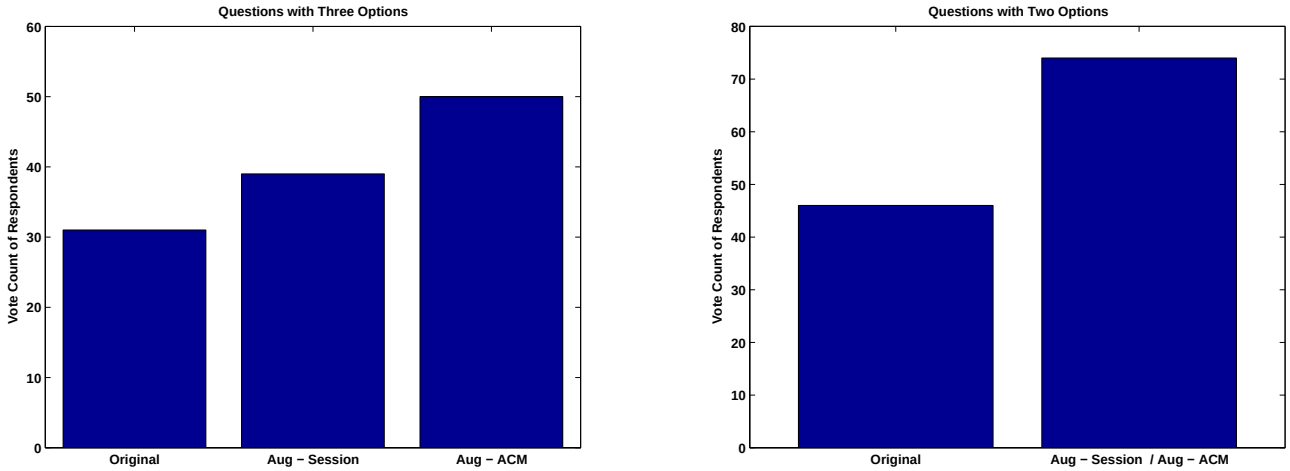

Figure 4: *User Study on Augmenting Titles*

metadata from ACM 2012 CCS is preferred. Once the survey was completed, we revealed the source of paper titles to the participants and presented the results. The researchers agreed that ACM 2012 CCS is preferred for augmenting paper titles even if the papers are not part of an ACM conference since they bring uniformity in metadata while session names can vary.

Prescription for Organizing Conference Papers in Digital Libraries

A research paper's worth to future readers depends on its content. However, the title of a paper helps it to be quickly discovered and hence must be meaningful. Augmenting paper titles particularly helps papers which have (a) very short titles, (b) long catchy titles but not directly relevant to the content of the paper.

To augment titles, one can use metadata from paper submission like keywords and topics of interest; presentation like session names, and publication like ACM 2012 CCS. Usually, submission metadata is captured in conference management systems but lost thereafter. Session names are particular to each conference and can become inconsistent over time. A classification scheme like CCS offers stability across conferences and is thus preferred. In our limited user study, participants (researchers) prefered CCS over session names.

We now use the insights to prescribe how paper titles should be organized. If the paper appears in an ACM conference, its CCS classification is available during submission and publication. We suggest that available CCS should directly be added to titles of papers by indexing organizations to improve discovery. But for non-ACM conferences, which includes a wide variety of AI papers, one may not have paper submission metadata and only be left with session names. Fortunately, one can map session names to CCS and use the corresponding closest CCS labels to effectively augment titles.

Conclusion

Research papers are written to be discovered, read and cited. It is well known that the title of an article impacts how well it is discovered by potential readers. In this paper, we explored characteristics of titles of AI papers and that of highly cited papers and compared them with recommended guidelines. We found that although good and well cited paper titles have word length in narrow range that promotes recall, the AI papers can have a wide diversity. To improve title quality and make them more amenable for search, we investigated automatic augmentation of titles using session names of the conferences in which they appeared and well as from ACM classification. We experimented with a user group and found that they overwhelmingly prefer the augmented titles over the originals for being more helpful with search.

References

AAAI. 2012. Aaai-12 schedule. http://www.aaai.org/Conferences/AAAI/2012/aaai12schedule.pdf.

AAAI. 2013. Aaai-13 schedule. http://www.easychair.org/smart-program/AAAI-13/index.html.

ACM. 2012. 2012 acm computing classification system. In http://dl.acm.org/ccs_flat.cfm.

Anthony, L. 2001. Characteristic features of research article titles in computer science. In *IEEE Transactions on Professional Communication, Vol. 44, pp.187-194*.

Chapman, S. 2009. Simmetrics. http://www.dcs.shef.ac.uk/~sam/stringmetrics.html.

Cohen, W.; Ravikumar, P.; and Fienberg, S. 2003. A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, volume 3.

Falagas, M. E.; Zarkali, A.; Karageorgopoulos, D. E.; V, V. B.; and Mavrose, M. N. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. In *PLoS One, Vol.8(2), p.e49476*.

Habibzadeh, F., and Yadollahie, M. 2010. Are shorter article titles more attractive for citations? cross-sectional study of 22 scientific journals. In *Croatian Medical Journal, Vol. 51, No. 2, pp. 165170.*

Jamali, H. R., and Nikzad, M. 2011. Article title type and its relation with the number of downloads and citations. In *Scientometrics, online first.* DOI: 10.1007/s11192-011-0412-z.

Kumar, M. J. 2013. Making your research paper discoverable: Title plays the winning trick. http://mamidala.wordpress.com/2013/08/11/making-your-research-paper-discoverable-title-plays-the-winning-trick. Accessed 13 Aug 2014.

Lopez, C.; Prince, V.; and Roche, M. 2014. How can catchy titles be generated without loss of informativeness? In *Expert Systems with Applications,* 41 (4) : 1051-1062.

Mack, C. 2012. How to write a good scientific paper: title, abstract, and keywords. In *Journal of Micro/Nanolithography, MEMS, and MOEMS,* Vol.11(2), pp.020101-1 020101-4.

Sagi, I., and Yechiam, E. 2008. Amusing titles in scientific journals and article citation. In *Journal of Information Science,* Vol. 34, No. 5, pp. 680687.

Srivastava, B.; Ponnalagu, K.; Narendra, N.; and Kannan, K. 2007. Enhancing asset search and retrieval in a services repository using consumption contexts. In *Services Computing,* 2007. SCC 2007. IEEE International Conference on, 316–323.

USC. 2014. Making your research paper discoverable: Title plays the winning trick. In *Univ. of Southern California's paper guideline,* At http://libguides.usc.edu/content.php?pid=83009\&sid=801403. Updated 4 Aug 2014.

WWW. 2012. Www-12 schedule. http://www2012.org/wwwprogram/getinfo.php?type=technical.

WWW. 2013. Www-13 schedule. http://www2013.org/program.