Proceedings of

BIS 2007 Workshops

on

Social Aspects of the Web (SAW 2007)

Ontology Evolution (OnE 2007)

Web Services Interactions, Quality and SLAs (WS-IQS 2007)

Dominik Flejter, Monika Kaczmarek, Marek Kowalkiewicz,
Peter Plessers, Krzysztof Węcel, Dominik Zyskowski (Eds.)

ISBN-10: 83-916842-4-5
ISBN-13: 978-83-916842-4-5

Contents

Workshop on

Social Aspects of the Web (SAW 2007)

27th of April 2007,
Poznań, Poland

http://www.integror.net/saw/

Workshop Program Committee

John Breslin, DERI, NUI Galway, Ireland
Dominik Flejter, Poznan University of Economics, Poland
Marek Kowalkiewicz, SAP Research Brisbane, Australia
Marcin Paprzycki, Warsaw School of Social Psychology, Poland
Maarten de Rijke, University of Amsterdam, the Netherlands
Marcin Sydow, Polish-Japanese Institute of Information Technology, Poland

HOMOGENEOUS TEMPORAL ACTIVITY PATTERNS IN A LARGE ONLINE COMMUNICATION SPACE

Andreas Kaltenbrunner[*,‡,§], Vicenç Gómez[*,‡], Ayman Moghnieh[*], Rodrigo Meza[*,‡], Josep Blat[*,‡] and Vicente López[*,‡]

[*]*Universitat Pompeu Fabra, Departament de Tecnologia, Passeig de Circumval·lació 8, 08003 Barcelona, Spain*

[‡]*Barcelona Media Centre d'Innovació, Ocata 1, 08003 Barcelona, Spain*

ABSTRACT

The many-to-many social communication activity on the popular technology-news website Slashdot has been studied. We have concentrated on the dynamics of message production without considering semantic relations and have found regular temporal patterns in the reaction time of the community to a news-post as well as in single user behavior. The statistics of these activities follow log-normal distributions. Daily and weekly oscillatory cycles, which cause slight variations of this simple behavior, are identified. The findings are remarkable since the distribution of the number of comments per users, which is also analyzed, indicates a great amount of heterogeneity in the community. The reader may find surprising that only two parameters, those of the log-normal law, allow a detailed description, or even prediction, of social many-to-many information exchange in this kind of popular public spaces.

KEYWORDS

Social interaction, information diffusion, log-normal activity, heavy tails, Slashdot

1. INTRODUCTION

Nowadays, an important part of human activity leaves electronic traces in form of server logs, e-mails, loan registers, credit card transactions, blogs, etc. This huge amount of generated data allows to observe human behavior and communication patterns at nearly no cost on a scale and dimension which would have been impossible some decades ago. A considerable number of studies have emerged in recent years using some part of these data to investigate the time patterns of human activity. The studied temporal events are rather diverse and reach from directory listings and file transfers (FTP requests) (Paxson and Floyd 1995), job submissions on a supercomputer (Kleban and Clearwater 2003), arrival times of consecutive printing-job submissions

[§]Corresponding author: e-mail: andreas.kaltenbrunner@upf.edu; Fax: +34 93 542 2517

(Harder and Paczuski 2006) over trades in bond (Mainardi et al. 2000) or currency futures (Masoliver et al. 2003) to messages in Internet chat systems (Dewes et al. 2003), online games (Henderson and Bhatti 2001), page downloads on a news site (Dezso et al. 2006) and e-mails (Johansen 2004). A common characteristic of these studies is that the observed probability distributions for the waiting or inter-event times are heavy tailed. In other words, if the response time ever exceeds a large value, then it is likely to exceed any larger value as well (Sigman 1999). A recent study (Barabási 2005) tries to explain this behavior under the assumption that these heavy tailed distributions can be well approximated by a power-law or at least by a power-law with an exponential cut-off (Newman 2005). The cited study presents a model which seems to explain the distribution of e-mail response times and has been used later to account for the inter-event times of web-browsing, library loans, trade transactions and correspondence patterns of letters (Vázquez et al. 2006). However, the hypothesis of a power-law distribution is not generally accepted, at least in case of e-mail response times. Stouffer et al. (2006) claim that the data can be much better fitted with a log-normal distribution (Limpert et al. 2001). This debate has been repeated across many areas of science for decades, as noticed by Mitzenbacher (2004).

To the authors' knowledge no study of this type has been performed on systems where social interaction occurs in a more complex manner than just person to person (one-to-one) communication. We think it is valuable to analyze the temporal patterns of the many-to-many social interaction on a technology-related news-website which supports user participation. We have chosen Slashdot[1], a popular website for people interested in reading and discussing about technology and its ramifications. It gave name to the "Slashdot effect" (Adler 1999), a huge influx of traffic to a hosted link during a short period of time, causing it to slow down or even to temporarily collapse.

Slashdot was created at the end of 1997 and has ever since metamorphosed into a website that hosts a large interactive community capable of influencing public perceptions and awareness on the topics addressed. Its role can be metaphorically compared to that of commercial malls in developed markets, or hubs in intricate large networks. The site's interaction consists of short-story **posts** that often carry fresh news and links to sources of information with more details. These posts incite many readers to **comment** on them and provoke discussions that may trail for hours or even days. Most of the commentators register and comment under their nicknames, although a considerable amount participates anonymously.

Although Slashdot allows users to express their opinion freely, moderation and meta-moderation mechanisms are employed to judge comments and enable readers to filter them by quality. The moderation system was analyzed by Lampe and Resnick (2004) who concluded that it upholds the quality of discussions by discouraging spam and offending comments, marking a difference between Slashdot and regular discussion forums. This high quality social interaction has prompted several socio-analytical studies about Slashdot. Poor (2005) and Baoill (2000) have both conducted independent inquiries on the extent to which the site represents an online public sphere as defined by Habermas (1962/1989).

Given that a great amount of users with different interests and motivations participate in the discussions, one would expect to observe a high degree of heterogeneity on a site like Slashdot. However, what if the posts and comments were analyzed just as imprints of an occurring information exchange, with no regard to semantic aspects? Is there a homogeneous behavior pattern underlying heterogeneity? To answer these and related questions we collected and studied one

[1] http://www.slashdot.org

year's worth of interchanged messages along with the associated metadata from Slashdot. We show here that the temporal patterns of the comments provoked by a post are very similar, indicating that homogeneity is the rule not the exception. The temporal patterns of the social activity fit accurately log-normal distributions, thus giving empirical evidence of our hypothesis and establishing a link with previous studies where social interaction occurs in a simpler way.

Finally, our analysis allows more insight into questions such as: is there a time-scale common to all discussions, or are they scale-free? What does incite a user to write a comment, is it the relevance of the topic, or maybe just the hour of the day? Can we predict the amount of activity triggered by a post already some minutes after it has been written? Which type of applications can we devise on the basis of using these conclusions?

The rest of the article is organized as follows: In section 2 we briefly explain the process of data acquisition. We then present the results in section 3 providing first an overview of the global activity and then explaining our analysis in detail. We finish the paper with section 4 where we discuss the results.

2. METHODS

In this section we explain the methods used to crawl and analyze Slashdot. The crawled[2] data correspond to posts and comments published between August 26th, 2005 and August 31th, 2006. We divided the crawling process into two stages. The first stage included crawling the main HTML (posts) and first level comments and the second stage covered all additional comment pages. Crawling all the data took 4.5 days and produced approximately 4.54 GB of data. Post-processing caused by the presence of duplicated comments was necessary (due to an error of representation on the website). Although a high amount of information was extracted from the raw HTML (sub-domains, title, topics, hierarchical relations between comments) we concentrated only on a minimal amount of information: **type** of contribution (either post or comment), its **identifier**, **author**'s identifier and **time-stamp** or date of publishing. The selected information was extracted to XML-files and imported into Matlab where the statistical analysis was performed. Table 1 shows the main quantities of the crawling and the extracted data.

Table 1. Main quantities of crawling and retrieved data.

Period covered	26-8-05 − 31-9-06
Time needed for crawling	4.5 days
Amount of data mined	4.54 GB
Posts	10016
Comments	2075085
Commentators	93636
Anonymous comments	18.6%

The time-stamps of post and comments can be obtained from Slashdot with minute-precision and corresponded to the EDT time zone (= GMT−4 hours). They allow to calculate the follow-

[2]Software used: wget, Perl scripts, and Tidy on a GNU/Linux, Ubuntu 6.0.6 OS.

ing two quantities:

The **Post-Comment-Interval (PCI)** stands for the difference between the time-stamps of a comment and its corresponding post.

The **Inter-Comment-Interval (ICI)** refers to the difference between the time-stamps of two consecutive comments of the same user (no matter what post he/she comments on).

3. RESULTS

In this section we first give an overview of the global activity looking at the data on different temporal scales and analyzing some relations between variables of interest. We then focus on the activity provoked by single posts and analyze the behavior of single users, concentrating on the most active ones.

3.1 Global cyclic activity

As previously explained, comments can be considered as reactions triggered by the publishing of posts. This difference in nature between both types of contributions justifies a separate analysis of their dynamics.

Figure 1 shows (normalized) mean activity and standard deviations of both posts and comments. It illustrates patterns in agreement with the social activity outside the public sphere. Figure 1a shows regular, steady activity during working days which slows down during weekends. This weekly cycle is interleaved by daily oscillations illustrated in Figure 1b. The daily activity cycle reaches its maximum at 1pm approximately and its minimum during the night between 3am and 4am. Although Slashdot is open to public access around the world, we see that its activity profile is clearly biased towards the American time-schedule.

Interestingly, although post activity shows more fluctuations and higher standard deviations than comment activity, there is little discrepancy between their mean temporal profiles. This difference in the deviations is not surprising given the greater number of comments (see Table 1). We notice that the standard deviations of the daily post- and commenting activities also show

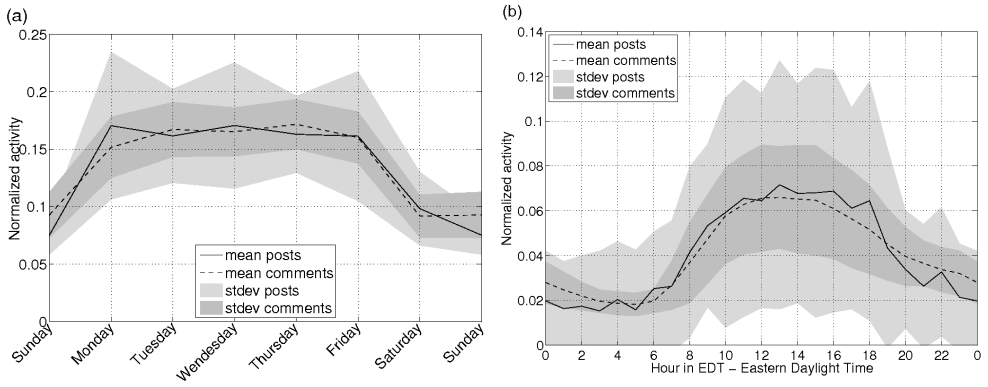

Figure 1. **(a)** Weekly and **(b)** daily activity cycles.

similar cyclic behavior (Figure 1b).

3.2 Post-induced activity

In this section we analyze the activity (comments) a post induces on the site. The histogram of Figure 2 gives an idea of the number of comments the posts receive. Note that half of the posts provoke more than 160 comments and some of them even trigger more than 1000. To analyze the time-distribution of these comments we study their post-comment intervals (PCIs).

3.2.1 Analysis of the activity generated by a single post

We are especially interested in the resulting probability distribution of all the PCIs of a certain post. This distribution reveals us the probability for a post to receive a comment t minutes after it has been published. Figures 3a and 3b show this distribution for a post which provoked 1341 comments. Although there are some important fluctuations, the characteristic shape of the probability density function (pdf) resembles a log-normal distribution. This becomes even clearer if the cumulative probability distribution (cdf) is observed, since there the fluctuations of the pdf are averaged out. Figures 3c and 3d show a good fit of the PCI-cdf of the data with the cdf of the log-normal distribution.

To classify the quality of the fit we have used a normalized error measure ε based on the ℓ^1-norm (see Appendix A). For the post shown in Figure 3 we obtain $\varepsilon = 0.007$, meaning that the average error is below 1%.

The PCI-cdf of three more posts can be observed in Figure 4. The top two sub-figures show good fits, indicating that the PCI is well approximated even for a small number of comments. However, the fit is not that accurate for all posts. For example, the comments of the post shown in Figure 4 (bottom) start to show considerable different behavior from the expected log-normal approximation about 3 hours after its publication. The activity is lower than the predicted one, but starts to increase again at about 6am in the morning the next day. At around at 8:30pm it increases further to recover the lost activity during the night. More such increases and decreases of activity can be observed during the following days. The time-spans of variations in activity coincide quite exactly with the average daily activity cycle shown in Figure 1b. We analyze this coincidence further in the next section.

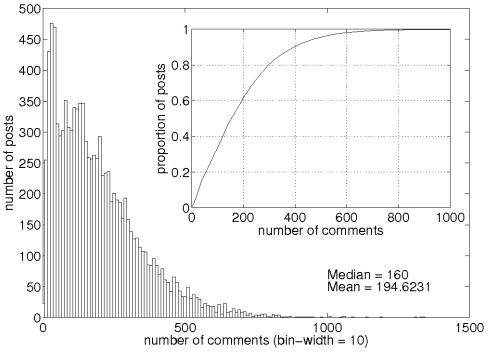

Figure 2. Histogram of the number of comments per post (inset shows the corresponding cdf).

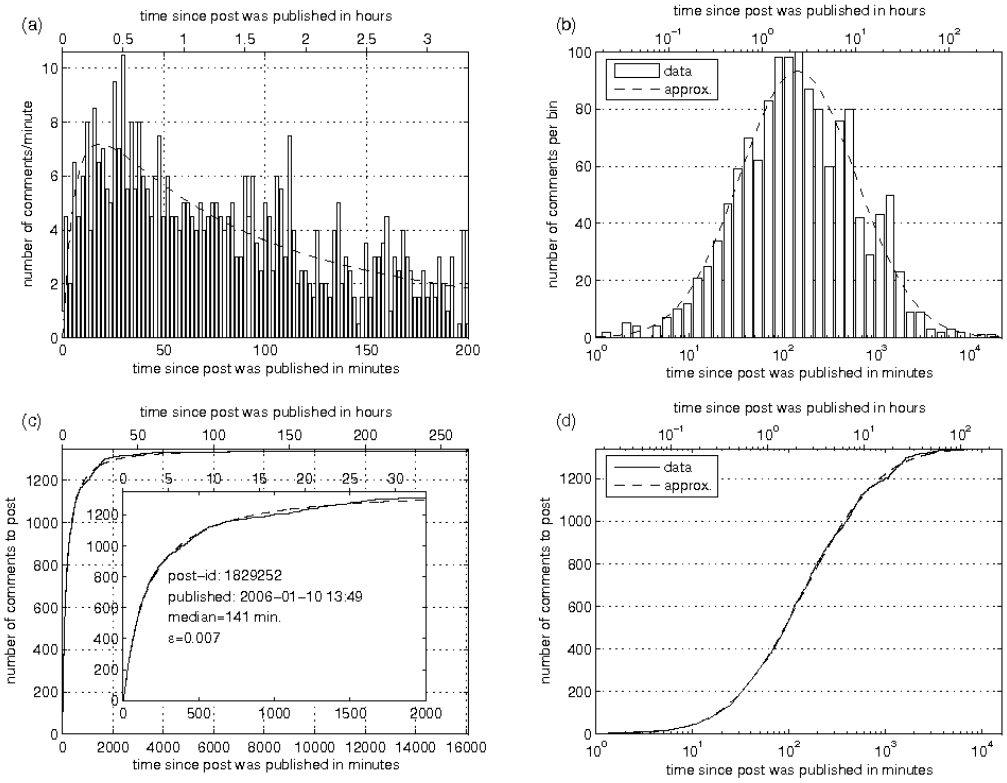

Figure 3. Log-normal approximation (dashed lines) of the PCI-distribution (solid lines and bars) of a post which received 1341 comments. **(a)** Comments per minutes (bin-with= 2 for better visualization) for the first 200 minutes after the post has been published. **(b)** Same as (a) in logarithmic scale. **(c)** The cumulative distribution of the data shown in (a). Inset shows a zoom on the first 2000 minutes. **(d)** Same as (c) in logarithmic scale.

3.2.2 Comparison of posts

With the log-normal shape of the PCI-distribution identified, we focus on the quality of this approximation in general. We therefore calculate the error measure ε of the fit for all posts which received comments. The resulting distribution of ε can be seen in Figure 5a. For 87% of the posts the approximation error ε is lower than 0.05, and for 29% lower than 0.02.

If we take a closer look at the data, we notice a dependence of ε on the publishing-hour of a post (Figure 5b). The best fit is reached when the post is published between 6am and 11am. Then the mean error increases successively until 11pm to stay high during the night and recover again in the early morning.

This behavior can be understood looking at the daily activity cycle (Figure 1b). The less time the community has to comment on a post during the time-window of high activity, the greater is the need to comment on it the next time the high activity phase is reached, and hence the expected log-normal behavior is altered. Figure 4 (bottom) gives examples of such a late post (published at 10:35pm).

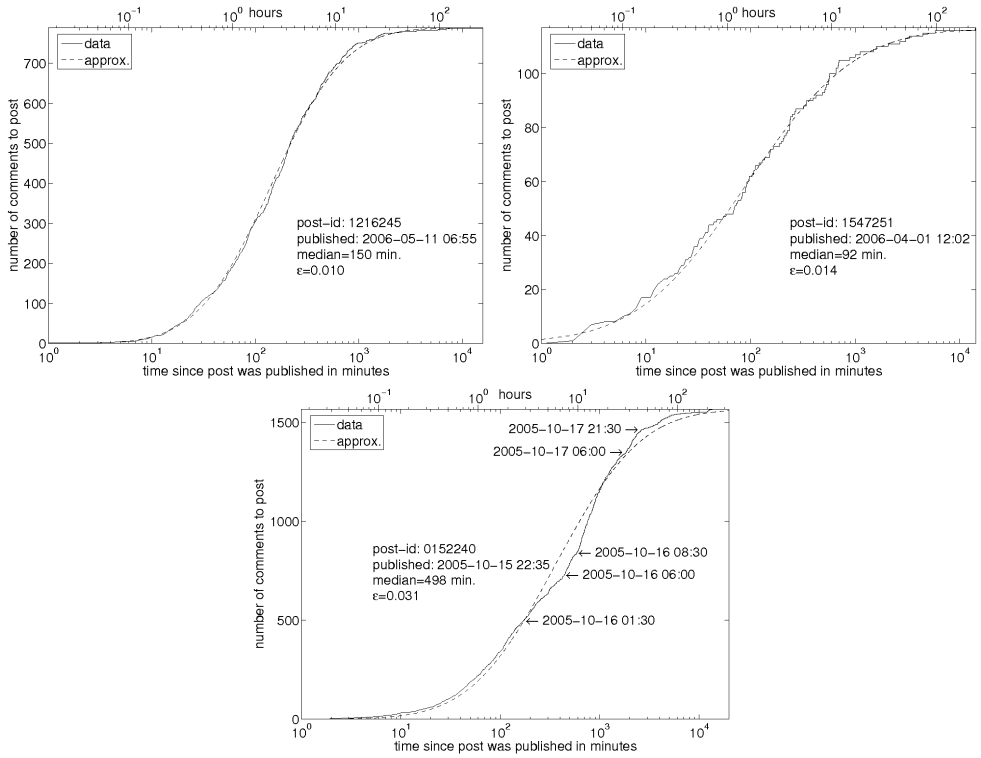

Figure 4. Log-normal approximation of the PCI-distribution of 3 different posts.

The good quality of the approximation allows us to describe the activity triggered by a post with only two parameters, the median[3] and the geometric standard deviation σ_g of the PCI-pdf, commonly used to compare log-normally distributed quantities (Limpert et al. 2001). Figure 6 shows the distribution of these quantities. The inset shows σ_g, which is centered around 1.036 and very similar for all posts. The median of the post-induced activity on the other hand shows more variations, but is rather short (for 50% of the posts it is below 2.5 hours, for 90% below 6.5 hours) compared to the maximum PCI (approx. 12 days). We can thus conclude that although the total activity a post generates covers a large time interval the major part of the activity happens within the first few hours after the post's publication.

3.3 User dynamics

In this section we analyze the activity on Slashdot taking the authorship of the comments into account. We first study the distribution of activity among all the users participating in the debates and then focus on the temporal activity patterns of single users.

[3]Note that the median coincides with the geometric mean for a log-normally distributed random variable.

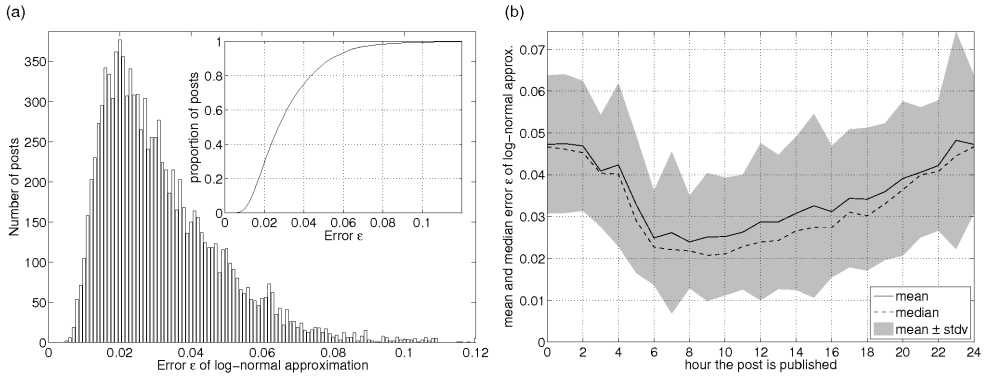

Figure 5. **(a)** Errors ε of the log-normal approximation of the PCI-cdf (bin-width $= 10^{-3}$). Inset shows the corresponding cdf. **(b)** Dependence of mean and median of the approximation error ε on the hour the post is published.

3.3.1 Global user activity

The activity of all users is best illustrated by the distribution of the number of comments per user. It is shown in double-logarithmic scale in Figure 7a. The obtained distribution follows quite closely a straight line, suggesting a power-law probability distribution governing this relation. We note that 53% of the users write 3 or less comments whereas only 93 users (0.1%) write more than 1000 comments. Indeed, after applying linear regression as in other studies (Faloutsos et al. 1999, Albert et al. 1999) we obtain a quite large correlation coefficient $R^2 = -0.97$ for an exponent of $\gamma = -1.79$.

However, if we apply rigorous statistical analysis as proposed in Goldstein et al. (2004) the picture changes. First, we estimate the power-law exponent computing the less biased maximum likelihood estimator (MLE). The resulting exponent $\gamma = -1.5$ differs significantly from the previous one and is illustrated in Figure 7 (dashed-line). Although Figure 7a tempts one to accept the power-law hypothesis, the cdf shown in Figure 7b discards it. It is thus not surprising that the Kolmogorov-Smirnov test forces us to reject the power-law hypothesis with statistical

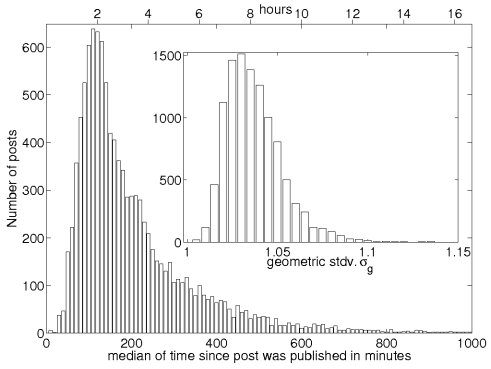

Figure 6. Histograms of medians (bin-width $= 10$) and geometric standard deviations (inset, bin-width $= 0.005$) of the PCI-distributions.

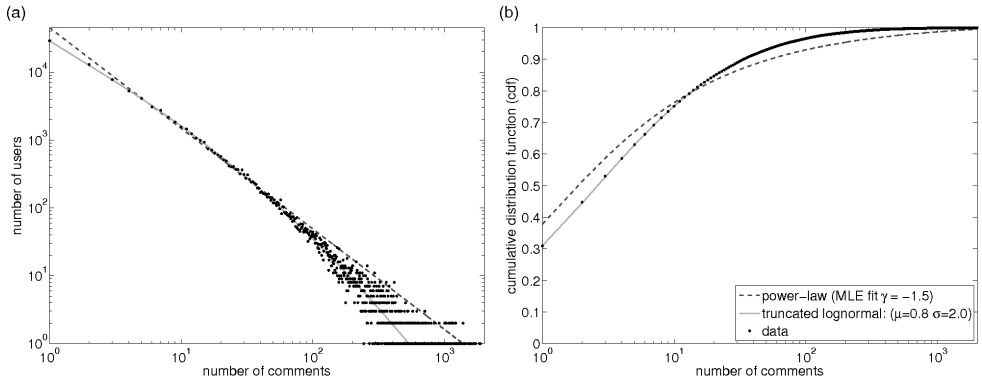

Figure 7. **(a)** Histogram of the number of comments per user and **(b)** and its corresponding cdf.

significance at the 0.1% level.

As an alternative hypothesis to describe the data we propose a (truncated) log-normal probability distribution, shown in Figure 7 as grey-solid-line. Its parameters are found using the MLE. Clearly, the fit is better using this hypothesis. We remark that in many studies some data points (considered outliers) are discarded to improve the power-law fit. Here, in contrast, the truncated log-normal approximation can characterize the entire data-set.

3.3.2 Single user dynamics

After characterizing the user activity at a general level, we investigate the temporal behavior patterns of single users . The analysis concentrates on the two most active users (to protect their privacy we call them user1 and user2). Table 2 shows the number of commented posts and the total number of comments these two users published during the time-span covered by our data.

Table 2. Contributions of the two most active users.

	user1	user2
commented posts	1189	1306
comments	3642	3350

We focus on the distribution of the PCIs of all of their comments as well as on their inter-comment-interval (ICI) distribution, i.e. the time-difference between two comments of the same user.

The PCI-cdf (see Figure 8a) of the two users can also be approximated by a log-normal distribution, although the fit is worse than in the case of the post-induced comment activity. Again we notice a clear dependence of the quality of the fit on the activity cycle (shown in the insets of Figure 8a). The approximation is much better for user1, whose daily and especially weekly activity cycles are much more balanced than those of user2. The activity of the latter user concentrates almost exclusively on the working hours from Monday to Friday. Hence his PCI-distribution shows a clear decrease after 8 but increases again after 16 hours. This increase is less pronounced if only the first comment to a post is considered (data not shown), indicating that the user frequently rechecks the posts he commented the day before to participate again in an ongoing discussion.

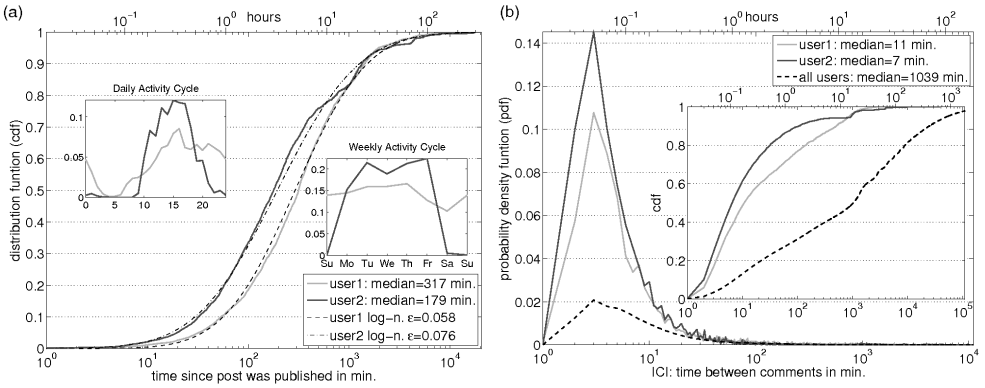

Figure 8. Activity patterns of the two most active users: **(a)** PCI-distributions, insets shows daily and weekly activity cycles. **(b)** Distribution of the inter-comment intervals (ICI) compared with the whole population (dashed line).

The same effect can be observed in their ICIs, which are illustrated in Figure 8b. There the cdf (inset of Figure 8b) of user1 shows an even more pronounced increase around an ICI of 16 hours. We further observe that the ICI-pdf peaks for both users as well as for the whole population at 3 minutes. This is probably caused by an anti-troll filter (Malda 2002), which should prevent a user from commenting more than once within 120 seconds. The medians of the ICI-distributions of user1 and user2 are rather short (11 and 7 minutes respectively) compared to the median of the whole population (about 17 hours), indicating that the two users engage in discussions frequently during their activity phase.

4. DISCUSSION

The special architecture of the technology-related news website Slashdot allowed us to analyze the temporal communication patterns of an online society without considering semantic aspects. The site activity is driven by news-posts which provoke communication activity in the form of comments.

Despite the great amount of users participating in the discussions, close to 10^5 in the data we have studied, and the diversity of themes (games, politics, science, books,etc.) some simple patterns can be identified, which repeat themselves over and over again. One of these patterns appears in the shape of the distribution of time differences between a post and its comments (the PCIs). It can be well approximated by a log-normal distribution (Figures 3 and 4) for most of the posts. The only remarkable deviations from these approximations are caused by oscillatory daily and weekly activity patterns (Figure 1), which become less noticeable if a post is published early in the morning (Figure 5a).

In single user behavior an akin pattern appears in the PCI-distribution of all of the comments a user writes to several posts (Figure 8a). Again deviations are caused by the circadian cycle. Another interesting pattern can be observed analyzing the ICI of single-users, i.e. the time-span between two consecutive comments of a certain user. In the case of the two most active users (Figure 8b) the ICI-distributions are very similar, which further supports our hypothesis of the

existence of homogeneous temporal patterns on Slashdot.

We would expect that the time-spans between publishing and reading of a post also follow a log-normal pattern. This could be easily verified checking the server logs of Slashdot or access-times of an external homepage linked by a Slashdot post. Such a study has been performed to show the Slashdot effect (Adler 1999), but the scale of the data presented does not allow to draw significant conclusions. Further investigation is needed to verify this claim.

Log-normal temporal patterns similar to those described above were found in person-to-person communication by Stouffer et al. (2006), who investigated the waiting and inter-event times of an e-mail activity dataset. A second coincidence between their study and our findings is that the number of comments (or e-mails in their case) can be well approximated by the same distribution (a truncated log-normal in this case). The temporal patterns of the e-mail data were previously claimed to show power-law behavior, which would be explained by a queuing model (Barabási 2005). Although this model might allow insight into other types of human activity (Vázquez et al. 2006) it is not able to account for the observed log-normal behavior patterns. We hope therefore to encourage further research towards a theoretical understanding of the underlying phenomena responsible for this apparently quite general human behavior pattern.

Our results indicate that communication activity on Slashdot can be described using only two parameters, i.e. the median and the geometric standard deviation (Figure 6). The medians are very low compared to the overall duration of the activity provoked by a post. Although the posts might be available for commenting during more than 10 days, the first few hours decide whether they will become highly debated or just receive some sporadic comments. We would therefore expect that the simplicity of the approximation together with the high initial activity should make an accurate prediction of the expected user behavior feasible at an early phase after a post has been put online. The accuracy of such forecasting is subject of current research and will be published elsewhere.

An early characterization of the activity triggered by a post could be applied, for instance, on dynamic pricing or placing of online advertisements or on the improvement of online marketing. The success of a campaign might be predicted already after a short time-period, thus allowing an early adaptation of the strategy of information diffusion. In this context the viral marketing concept (Leskovec et al. 2006) which relies on personal communication might be the most promising field.

In our opinion, the regular communication activity patterns described in this work may be relevant in two aspects. The first, simpler one, is related to applications where a better understanding of information trade in the web translates easily into a better description, and even quantification, of Internet audience. But a second, more complex, aspect is related to the human "communicative" behavior uncovered at present time: Internet based communication capabilities. We face a new, large scale, all-to-all public space in which a novel kind of social behavior arises, a scenario that we do not yet fully understand. However, we should not forget that the new activity is being largely recorded and the data can be available for research. The work presented in this contribution is a good example of how those data can be collected and analyzed to give, at least, a quantitative description of the behavior. This is a first step towards a more ambitious target: to develop "ab initio" models for the population dynamics of message interchange, which is also the goal of our current research.

ACKNOWLEDGMENTS

This work has been partially funded by Càtedra Telefónica de Producció Multimédia de la Universitat Pompeu Fabra.

REFERENCES

Adler, S., 1999. The Slashdot Effect, an analysis of three Internet publications. Published online.

Albert, R. et al, 1999. The diameter of the world wide web. *Nature* **401**:130.

Baoill, A. Ó., 2000. Slashdot and the Public Sphere. *First Monday* **5**(9).

Barabási, A. L., 2005. The origin of bursts and heavy tails in human dynamics. *Nature* **435**:207–211.

Dewes, C. et al, 2003. An analysis of Internet chat systems. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pp. 51–64, New York, NY, USA. ACM Press.

Dezso, Z. et al, 2006. Dynamics of information access on the web. *Physical Review E* **73**:066132.

Faloutsos, M. et al, 1999. On Power-law Relationships of the Internet Topology. In *SIGCOMM*, pp. 251–262.

Goldstein, M. L. et al, 2004. Problems with fitting to the power-law distribution. *The European Physical Journal B* **41**(2):255–258.

Habermas, J., 1962/1989. *The Structural Transformation of the Public Sphere : Inquiry into a Category of Bourgeois Society*. Cambridge, MA: MIT Press.

Harder, U. and Paczuski, M., 2006. Correlated dynamics in human printing behavior. *Physica A* **361**:329–336.

Henderson, T. and Bhatti, S., 2001. Modelling user behaviour in networked games. In *MULTIMEDIA '01: Proceedings of the 9th ACM International Conference on Multimedia*, pp. 212–220, New York, NY, USA. ACM Press.

Johansen, A., 2004. Probing Human Response Times. *PHYSICA A* **338**:286.

Kleban, S. D. and Clearwater, S. H., 2003. Hierarchical Dynamics, Interarrival Times, and Performance. In *SC '03: Proceedings of the 2003 ACM/IEEE conference on Supercomputing*, p. 28, Washington, DC, USA. IEEE Computer Society.

Lampe, C. and Resnick, P., 2004. Slash(dot) and burn: Distributed Moderation in a Large Online Conversation Space. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 543–550, New York, NY, USA. ACM Press.

Leskovec, J. et al, 2006. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pp. 228–237, New York, NY, USA. ACM Press.

Limpert, E. et al, 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience* **51**:341–352.

Mainardi, F. et al, 2000. Fractional calculus and continuous-time finance II: the waiting-time distribution. *Physica A* **287**:468–481.

Malda, R., 2002. Slashdot FAQ: Comments and Moderation. http://slashdot.org/faq/com-mod.shtml#cm2000.

Masoliver, J. et al, 2003. Continuous-time random-walk model for financial distributions. *Physical Review E* **67**:021112.

Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1**(2):226–251.

Newman, M. E. J., 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46**:323–351.

Paxson, V. and Floyd, S., 1995. Wide area traffic: The failure of Poisson modeling. *IEEE-ACM Transactions On Networking* **3**:226–244.

Poor, N., 2005. Mechanisms of online public sphere: The web site Slashdot. *Journal of Computer-Mediated Communication* **10**(2).

Sigman, K., 1999. Appendix: A primer on heavy-tailed distributions. *Queueing Systems* **33**:261–275.

Stouffer, D. B. et al, 2006. Log-normal statistics in e-mail communication patterns. e-print physics/0605027.

Vázquez, A. et al, 2006. Modeling bursts and heavy tails in human dynamics. *Physical Review E* **73**:036127.

APPENDIX A ERROR MEASURE ε

We use the following distance measure to calculate the error of log-normal approximation of the data. The distance between approximation and data is only calculated for the time-bins (i.e. minutes) where a post actually receives a comment to avoid a distortion of the error measure by the periods with low comment activity.

Definition 1. *Let $\mathbb{T}$ be the set of time-bins where a post receives at least one comment and T its cardinality. We define then the approximation error ε of a function $f(t)$ approximating $g(t)$ (both defined for all $t \in \mathbb{T}$) as the normalized ℓ^1-norm of $f(t) - g(t)$:*

$$\varepsilon = \sum_{t \in \mathbb{T}} \frac{|f(t) - g(t)|}{T} \tag{1}$$

If $f(t)$ and $g(t)$ are cumulative probability density functions (i.e. $0 \leq f(t) \leq 1$ and $0 \leq g(t) \leq 1$), it follows that $0 \leq \varepsilon \leq 1$.

A FRAMEWORK FOR EXPLORING RELATIONSHIPS BETWEEN ONLINE COMMUNITY CHARACTERISTICS AND REGULATION PRINCIPLES

Eleonore ten Thij
Justin de Nooijer
Fac. of Science, Information and Computing Sciences, Utrecht University
P.O box 80 089 NL-3508 TB Utrecht
E.tenThij@cs.uu.nl

ABSTRACT

In this paper, we explore a framework for researching relationships between community characteristics and regulation principles. Different regulation principles are supported by different website features. Ultimately, our goal is to help community operators to deepen the appreciation of their community sites by providing empirically validated insights which website features might support their online community best.
We first determined which community characteristics and regulation principles should be considered, based on a literature search. We then analyzed 31 Dutch and English, national and regional online newspaper communities.
Analysis showed some interesting relationships between individual community characteristics and regulation principles. The framework was able to discriminate between two types of community as well, on the basis of our data, but could not relate these types to (sets of) different regulation principles. We will therefore suggest some improvements of our framework.

KEYWORDS

online community characteristics; regulation principles; typology; design

1. INTRODUCTION

Who observes directories of online communities may notice that webspaces for online communities are created, populated, and abandoned on a regular base. Some online communities seem more or less sustainable, others do not manage to even get the critical mass to really get started. We can find (free) community software on the internet with features built in to empower community formation, like profiles to express personal identity and negotiate social identity, or rating and ranking systems to ensure the quality of member contributions, express roles or help building commitment. However, we do not really know that much yet about whether these features indeed succeed in empowering online communities.
Our research seeks to further the understanding of how the design of community sites may effect community formation. In a previous study we researched how categories of website features expressing success factors and guidelines found in the literature contributed to the appreciation of community sites. (Ten Thij, Van de Wijngaert, 2006). This study did not take into account yet, that these categories may take different values for different types of online communities. Different types of communities may develop different sets of regulation principles that need to be supported by different website features. In previous research, we found that members of gaming communities are more appreciative of being engaged in co-developing and maintaining the community site than members of consumer-to-consumer communities are (Ten Thij, 2007). Consequently, gaming communities may be more appreciative of website features allowing members to do so, for instance by means of an elaborated and refined system of privileges to support moderator functions.

Most typologies found in the literature, however, do not take into account that community characteristics may vary within different socio-cultural settings. For example, from observations and signals from the

newspaper branch in the Netherlands, we can state that the effort of setting up and maintaining an online community is a complicated task: they may attract fewer members than expected, they may show relatively low levels of interaction, or struggle with abuse (spamming, flaming, racial slur). In the literature on guidelines and design principles for online communities (self-)regulation (i.e. policies, rules that engage members in co-developing the community) is considered an important issue (Kollock, 1997; Kim, 2000; Preece, 2003). Newspapers may find it difficult to allow their community to self-organize because their reputation may be at stake, and their staff traditionally is more used to creating content than to supporting interaction. However, online newspaper communities in other countries may very well behave differently, due to different cultural norms and values with respect to community formation and newspaper policies. Likewise, online newspapers covering national or regional markets may also take different perspectives on setting up and managing online communities, since they might differ in how they relate to local communities.

In this paper, we will built a framework for assessing relationships between community characteristics and (self-)regulation principles. The aim of this framework is enabling future research to assess success factors differentiated for specific types of online communities in different socio-cultural settings. Our research questions therefore are:

- *how are community characteristics related to (self-)regulation principles?*
- *are community characteristics differently related to (self-)regulation principles in different socio-cultural settings?*

We will first elaborate on the framework, and then present and discuss the results of a first tentative test of the framework on online communities related to Dutch and British national and regional newspapers.

2. RESEARCH FRAMEWORK

2.1 Community Characteristics

Our starting point for identifying variables that may discriminate between different types of online communities was Porter (2004), who provides a state-of-the-art pre-defined typology, building upon and attempting to improve earlier typologies. On the first level her typology discerns member-initiated and organization-sponsored virtual communities. Member-initiated communities are characterized by either having a social or professional relationship orientation. Relationships within an organization-sponsored community occur between members amongst each other as well as between members and the organization, and can be of a commercial, non-profit, or governmental nature. On a lower level she formulates a set of attributes to distinguish types of online communities empirically. In our category 'Community Characteristics' we use Porter's attributes, but we do not assume a pre-defined typology. Moreover, the variables are not always made operational in the same way:

- *Purpose*: can take the values 'relation' (*R*), 'entertainment' (*E*), 'action' (*A*), 'support' (*S*), or 'multiple' (*M*). We base these values on Ridings and Gefen 's (2004) research on motivation for participating in online communities and on Preece (2003). We counted and categorized news items on the community's front page, the highest score determining the value.
- *Place*: online (*O*) or hybrid (online and offline) (*H*). We based our scores here on signs of organized offline events on the website, and whether or not members discuss meeting offline.
- *Platform*: can take the values 'synchronous' (*S*), 'asynchronous' (*A*), or 'hybrid' (*H*). The value is determined by the presence of communication tools (chat, message board) on the website.
- *Population*: can take the value 'weak ties' (*O*) for interaction that does not show recurring usernames or apparent relationships. A value 'small group' (*S*) is given when a small number of re-occuring usernames and tight relations (i.e. enquiries about private life) are observed, and the community has fewer than 100 members. A value 'network' is scored when more loosely coupled relations are observed, while spam or flames occasionaly occur, and the community consists of 100 - 300 members. A value 'public' (*P*) is scored when a large number of usernames interact (in subgroups as well), while threads dedicated to flaming or spamming are observed, and the communitiy has over 300 members.

- *Outcome*: characterizes the 'gain' members get from participating in the community. Since we observed in previous research (Hoevers, Meulendijk, 2006) that member-initiated online communities can behave in very much the same way as profit-oriented organization-sponsored communities do, we choose to score the outcome for members as possible values (in stead of for community operators), since outcome may determine motivation to return to the online community. 'Outcome' is probably strongly related to 'Purpose', but they are not necessarily equivalent. The value 'relationships' (*R*) is scored when offline contact between members is initiated and encouraged. The value 'solutions' (*S*) is given when members support each other with solving problems (f.e. in support or auction communities). A value 'content' (*C*) is scored when members only discuss (news) items.

2.2 (Self-)regulation Principles

We define (self-)regulation principles as the policies and rules needed to manage the community's resources, and to generate commitment amongst its members. Kollock (1997) discusses heuristics, drawn from social sciences as well as from experience, that can help community developers to create a lively, elaborate social system. Partly, his 'design prinicples' were derived from Ostrom's (1990) work on non virtual communities ('commons'). Van Wendel de Joode (2005) researched open source communities on the implementation of Ostrom's principles. He grouped them together in 4 more usable clusters, namely *Boundaries*, *Creation of Commitment*, *Collective Choice*, *Appropriation and Provision*. In our framework we define the following categories of (self-)regulation principles:

- *Boundaries* will be characterized by two sub-variables, namely:
  Registration ($B_{registration}$): describes whether or not the community is protected by an entrance regulating system. A score of '1' means anyone can enter the community without registering. A score of '2' means the user is required to complete a short registration procedure (e.g. fill in his or her name, e-mail address and location). A score of '3' means entrance to the community is regulated by ways of an extensive profiling system, in which the users have to fill out many personal details (e.g. date of birth, address, occupation, religion, etc.).
  Specificity ($B_{specificity}$): measures the specificity of the community subject. A score of '0' means the community subject is very general, and therefore will not likely function as a 'natural boundary'; a score of '1' means the community subject is only interesting for a selected audience and will therefore very likely scare off potential intruders, hereby functioning as a 'natural boundary'.
- *Collective Choice* is measured in terms of who is controlling the development of the online community and the content offered on the community site. This variable has also two sub-variables:
  $C_{development}$: is measured firstly in terms of centralized and decentralized control (Walker & Dooley, 1999). Centralized control means a single control point (moderator) determines and dictates the rules and regulations. Decentralized control means multiple control points (community members) use their personal information on the community's state to determine applicable rules and regulations. A score of '1' means the control is centralized and users are not encouraged in any way to submit their opinions; this is the case when there is no notice in the community of users submitting their opinion. A score of '2' means the control is determined by a single control point, but accepting users' suggestions; this is the case when users are presented with the ability to submit suggestions by e-mail or a fill-out form. A score of '3' means the moderators base their decisions on members' input; this is the case when mechanisms such as a 'Community rules and regulations requests-section' on a forum, or a voting poll for the instalment or adjustment of rules and regulations are in place.
  $C_{content}$ measures whether or not users can post content themselves. A score of '1' will mean the automated offering and posting of content is not allowed (on the same level as for example an editor – on crucial pages of the community – though it is allowed in for example a forum or chat situation), a score of '2' will mean posting of content is allowed for only some users (e.g. those with higher rankings when a ranking system is in place, or those who are selected by the editors) and a score of '3' will mean posting is allowed for everyone, including unregistered members.
- *Appropriation and Provision* (Ap): characterizes to what extent rules of 'netiquette' are stated explicitly, and are being monitored, and to what extent rules are in place that (gradually) regulate the consuming of resources by the community members? A score of '1' means there are no explicit netiquette rules, and no formal rules implemented; this is the case when users can consume resources without the community

stimulating them to return the favour of making resources available. A score of '2' means there are few explicit netiquette rules, and some basic rules which cover the most basic aspects of community behaviour (e.g. controlling the amount of resources consumed), and they are brought to the attention of the user before he can consume the resource; this is the case when, for example, the user has the ability to consume only a certain amount of resources in a certain period of time. A score of '3' means there is an extensive explicit netiquette, and the ruling system is advanced and contains graduate appropriation; this is the case when specific groups of users are subject to specific sets of rules, for example moderators are bound by fewer rules than newly registered members.

- *Commitment (Co)*: Are there specific benefits for users which are aimed at provoking interaction or return visits? These benefits are not direct profit as discussed previously, but the 'extras' aimed at seducing members to revisit the community. In other words, what is offered by the community to its members in addition to the profit related to the community's purpose, in order to make it more interesting to engage in and continue interaction? A score of '1' means there are no benefits (other than the obvious interaction with like-minded people), '2' means there is basic functionality such as a news-letter or RSS feed, and '3' means there are advanced benefits, such as community-related (offline) events such as an excursion or just an organized meeting in a pub, or a chat-session with an expert .

2.3 Research Population

To analyze relationships between community characteristics and (self-)regulation principles within different socio-cultural contexts we randomly selected by means of a web search 31 Dutch and English online newspaper communities with both national and regional coverage (6 Dutch national (all national online newspapers), 9 Dutch regional, 7 English national and 9 regional online newspaper communities). The online communities differed in size and age within all groups.

Preliminary qualitative analysis showed that in this group there were no online communities primarily dedicated to entertainment or action. None of them allowed members to post content freely, other than on fora or chat rooms but some online newspapers allowed selected or higher ranked members to do so. Only one online newspaper community supported both synchronous and asynchonous communication, all others only supported asynchronous online communication.

3. RESULTS AND ANALYSIS

In this stage of developing the framework we did not yet formulate any specific hypotheses. We merely wanted to explore whether we would be able to find any significant relationships at all, thus testing the general applicability of the framework for discrimating between types of online communities within different socio-cultural settings.

Our informal "common sense" expectations were that regional online newspaper communities would show more small group relationships, since we felt that discussions would concern locally bounded interests, that would likely more directly affect members than subjects more related to (inter)national issues. Additionally, we expected that for the same reason members of regional online newspaper communities would more likely meet each other offline as well, and that relationships would more often be the 'gain' of participating in the online community. As a consequence, we expected that regional online newspaper communities would be more specific, and show less explicit rules of appropriation and provision. As far as differences between English and Dutch online newspapers are concerned, our informal expectations were that rules of appropriation and provision might be less explicit in Dutch online newspaper communities, since Dutch culture might be more oriented towards consensus building (Bakker, 2006).

After scoring we used Chi-Square test to calculate significant relationships ($p < 0.05$). Since we have only limited space here, we will present significant relationships only (see also Table 1, for results on chi-square tests):

- English online newspaper communities have a broader scope (multi-purpose), than the Dutch, that more often have just a singular purpose;

- Dutch online newspapers' communities have a more specific subject, which may serve as a natural boundary for visitors. This result also corresponds with that presented above. Dutch online newspaper communities may tend to aim at a specific target group (f.e. well-off singles or parents);
- English communities point out more explicitly and specifically which rules and behavioural norms their members and visitors have to comply with, and the consumption of content (e.g. the reading of articles, access to archives) is subject to a more advanced ruling system: there is a significant relation between country and the implementation of rules of appropriation and provision; The relationship between a community's purpose and whether or not posting of content (besides on a message board) is allowed (and if so, by whom) is significant as well. Only seven communities allow posting by certain types of members, four of which main purpose is information discussion, one is multi-purpose, one is relationship- , and one is support-oriented. The last two mentioned allow only content submission by registered members. All others do not allow submission of content whatsoever. So, the majority of information discussion and multi-purpose-related communities do not allow posting content. For information-related communities this might be explained by the newspapers' fear that members may post content that is less fact-based than news items written by professional journalists, thus threathening their reputation. This opposed to relationship communities, where one's submitted content is like an advertisement of his or her personality: submitting false or erroneous content in this case only affects the other members' opinion about the submitter, and whether or not they would want to engage in conversation and possibly a relationship with the advertised person;

Table 1 Chi-square tests Community characteristics and (self-)regulation principles[1]

Chi-Square test	*Pearson Chi square*	*df*	*Asymp. Sig. (2-sided)*	*Likelihood Ratio*	*Asymp. Sig. (2-sided)*
Country and Purpose	9.950	3	.019	11.179	.011
Country and Subject	4.045	1	.044	4.154	.042
Country and rules of Ap. and Prov.	7.306	2	.026	7.635	.022
Purpose and posting content by members	9.318	3	.025	9.018	.029
Population and Posting content by members	10.561	3	.014	9.521	.023
Purpose and Creation of Commitment	19.129	6	.004	11.177	.083
Outcome and creation of commitment	13.772	6	.032	14.697	.023
Outcome and rules of Ap. and Prov.	15.795	6	.015	18.910	.004
Coverage and rules of Ap. and Prov.	12.930	2	.002	15.696	.000
Coverage and registering for entrance	12.291	2	.002	15.995	.000

[1] N = 31

- The posting of content is also related to the size of a community's population. Smaller communities (small group and public) seldom allow posting of content, while 'no group' communities and large networks show a more diverse image. Because the data do not provide a clear insight as to what might cause this phenomenon, we are hesitant to draw conclusions on this point. Contrary to the findings, one

might expect the small-group communities to allow posting of content, for the mutual bonds most likely are tighter and trust could be less of an issue, as opposed to large communities where lots of members remain on the fringes, more or less anonymous. On the other hand, larger online newspaper communities may have more resources available to monitor the posting of content, and acquire and maintain the required software, and therefore allow posting more often;

- Table 1 shows a significant relation between a community's purpose and to what degree commitment from its members is stimulated by offering benefits. Information discussion communities tend to offer no extra benefits, while multi-purpose communities either do not offer any benefits, or offer extensive benefits such as expert chat sessions;
- *Creation of commitment* also shows a significant relation with *Outcome*. Generally, where the outcome consists of content, nothing is done in addition to the presentation of this content, to create commitment from members. The data also clearly show that communities whose outcome consists mainly of relationships and support, have more events aimed at binding members to the community;
- Table 1 shows a significant relation between outcome and the implementation of rules of appropriation and provision. The data show that communities where the outcome is generated content generally do not have any rules of appropriation and provision implemented. Relationship- and support-providing communities do have such systems implemented;
- *Coverage* has significant relations with the implementation of rules of appropriation and provision and members having to register for entrance to the community. The implementation of an appropriation and provision system is either not done or done to a moderate degree (scores of 1 or 2) in regional communities, whereas nationwide communities far more often have an advanced (score of 3) ruling system. Nationwide communities either allow everyone to enter, or request an extensive profile to be filled out upon registering; this latter request is not very common in regional communities.

So far our framework did show some interesting relationships between community characteristics and (self-)regulation principles. As a next step we tried to determine if the framework can indeed discriminate between types of community. We ran a Latent Class Analysis (LCA) that assumes that every cluster can be described by a chance distribution over the attributes *Purpose*, *Place*, *Population*, and *Profit*, while presupposing that these attributes are independant. We estimated models with different numbers of clusters, and it turned out that a model with two clusters gave the best BIC score (BIC (log-likelihood) = 262.10). A BIC score of a model M is calculated as follows:

BIC(M) = - 2*L(M) + npar(M) *log N, where L(M) is the value of the log-likelihood function under model M, evaluated in the maximum, npar(M) is the number of parameters, and N is the number of observations. The lower the BIC score, the better the model (Lazarsfeld, 1968, Vermunt, 1997).

Table 2 Cluster results from latent class analysis

Cluster 1 Information oriented	*Cluster 2 Multi-purpose*
AD	Volkskrant Parship
Metro	Daily Mail
NRC Handelsblad	Daily Mirror
Telegraaf	Daily Express
Trouw Moderne Manieren	Nieuws Op Urk
Daily Telegraph	Texelse Courant
Financial Times	The Argus
Sunday Mirror	Cambridge News
Guardian Unlimited	East Anglian Daily Times
De Stentor	Herts & Essex News
Leeuwarder Courant	Manchester Evening News
BN De Stem	The Cumberland
Brabants Dagblad	
Goors Nieuws	
Noordhollands Dagblad	
De Gooi- en Eemlander	
This Is London	
Daily Record	
Reading Evening	

Table 2 shows the clusters resulting from the LCA. We then tested wether clustering and variables were independent – wether the distribution of the variables over the clusters was the same for both clusters -, using Chi-square and Fisher exact tests. (see Table 3. *Platform* was excluded, since it scored the same on 30 papers).

Table 3 Contribution to clustering: Chi-square and Fisher exact test clusters and community characteristics variables

Variable	*Chi-square, sig.*	*Fisher exact, sig.*
Purpose	.0003	0
Place	.0000	0
Population	.9102	1
Profit	.0010	0

As we can see in Table 3 *Purpose* and *Place* contribute most to the clustering. *Population* hardly contributes to the clustering., which seems odd, considering that *Place* does. We will reflect on this later on.

From this we may conclude that the framework – on the basis of these data - can discriminate between a type of community that is information oriented, in which members meet each other online (Cluster 1) and a type of online community that is multi-purpose, in which members are not only interested in the information provided, but also form relationships, offer each other solutions to problems, and meet offline as well (Cluster 2).

However, these clusters were not confirmed when we performed LCA on (self-)regulation principles. Here a single cluster model gave the lowest BIC score (BIC (log-likelihood) = 285.65). We also did not find any significant differences between the individual or combined (self-)regulation principles for the two clusters (using independent t-tests). The framework may not contain the right categories to capture interesting differences in (self-)regulation, or the scoring itself may not have been flawless. In other words, the scoring method may not be sensitive and valid. Also our basic assumption that specific community characteristics relate to specific (self-)regulation principles may be false. On the basis of these data though, we must conclude that the framework is not fit yet to detect systematic relationships between types of online communities and different (self-)regulation principles.

4. DISCUSSION AND CONCLUSION

We have presented a framework for detecting relationships between online communities characteristics and (self-)regulation principles in different socio-cultural contexts, and explored its value by analyzing a number of Dutch and English national and regional online newspaper communities. The exploration resulted in some possibly interesting data and relationships, indicating answers to our research questions, which we summarize here:

- English online newspaper communities tend to have a multi-purpose function, whereas Dutch online newspaper communities serve a singular purpose. Next to that Dutch online newspaper communities tend to be more specific in their subjects. English online newspapers tend to offer a greater variety of services, like entertainment (playing games and watching video's), movie renting or dating, seem to partner with a number of commercial service companies, such as loan-offering or car-selling companies. They also offer a more extensive 'react-to-news items' functionality. This may be explained by different cultural norms towards independency of newspapers. Dutch newspapers might fear that partnerships with other commercial organizations would be regarded as endangering their objectivity, while English newspapers might feel less restricted in this respect. Additionally, it may be understood as a difference in perspective on what constitutes 'a third place' (Oldenburg, 1991). This should be researched though within a broader cultural and qualitative analysis;
- Possibly, because of this broader scope (multi-purpose function) and overall subject generality of English online newspaper communities, we found that English communities – far more often than Dutch communities – have implemented a more advanced appropriation and provision system of rules and

behavioural guidelines. Lacking a clearly perceivable boundary, they are more likely to attract a more heterogeneous group of members, in which it is more difficult to negotiate rules and norms informally. It may also confirm our expectation that Dutch online newspaper communities would be less explicit in stating rules of netiquette, since Dutch culture seems oriented towards consensus building (Bakker, 2006);

- The majority of online newspaper communities, especially information discussion and multi-purpose communities, and smaller communities, do not allow members to post content, and are also restricted in collective choice. This may be related to the afore mentioned difficulties newspapers may experience in allowing self-organization;
- Multi-purpose-, relationship- and support-oriented communities more often offer extra benefits to stimulate commitment than information discussion communities do. Possibly, the consumption of newly offered content itself is rewarding enough to make members return to the community. Multi-purpose-, relationship- and support-oriented communities organize more events. One can easily think of the benefits of such events for their members: relationship communities organizing offline meetings in local venues where singles can meet up, and support-providing communities offering the help of an expert in a chat session, etc.. Additionally, information discussion communities are comparatively more accessible in as far as they require less applyance with explicit rules of appropriation and provision. This makes sense, for the content in relationship and support communities can be far more privacy sensitive (consisting of extensive personal profiles including email addresses and pictures, or extensive descriptions of personal problems that are presented to members for the sake of obtaining a solution for a problem) than the content of an information discussion community (which mainly consists of opinions on news items). Thus, the consumption of the privacy sensitive information is (and probably should be) subject to more and more advanced rules;
- On the subject of access control, regional newspapers tend to have less constraints than national newspaper communities. Nationwide communities either allow everyone to enter, or request an extensive profile to be filled out upon registering, this latter request is not very common in regional communities. An explanation for this phenomenon can be that nationwide communities, asking for an extensive registration procedure also offer members access to archives, and may have relationships or support as (a) sub purpose(s), while regional newspapers do not. Our expectations that regional online newspapers would show more small-group relationships, more offline meetings, and fewer and less explicit rules, were not confirmed. This might mean regional online newspaper communities do not support existing local communities to a great extent;

Our approach, being explorative, still has some major weak points as well. Though it seems able to capture relationships between individual community characteristics and (self-)regulation principles, it is not able yet to relate types of online communities to (sets of) different (self-) regulation principles: we did find two different types of online community, information oriented and multi-purpose, but these types showed no systematic relationships with (sets of) different (self-)regulation principles.

We feel we can improve our framework by: (1) a better construction of variables: *Purpose* and *Boundary*, *specificity*, have been scored nominally, but would probably better be scored ordinally. This might accentuate the difference and relationship between both components; (2) a more advanced way of gathering data: several variables lend themselves better for data collection through member input by means of a questionnaire. *Purpose*, *Population*, *Outcome*, *Collective Choice*, and *Appropriation and Provision*, as far as informal rules are concerned, are good examples of this. Additionally, data on *Population* may be gathered by an automatic social network analysis of the postings contributed by members. This would also partly enable us (3) capturing the dynamics of online communities: online communities evolve constantly, are subject to experimentation, and quite often restricted in life span (even during the period this research was conducted, we have noticed (sections of) communities closing down due to abuse). It is also more than likely that communities have changed, evolved or shut down during the time that has passed since this research was conducted.

In previous research we found some empirical support for guidelines and design principles found in the literature in terms of appreciation factors, expressed in categories of website features (see Ten Thij & Van de Wijngaert, to appear). These categories of website features showed significant relationships with appreciation of online community sites. Once proven valid, this framework may be used to empirically assess relationships between community characteristics within different socio-cultural contexts, and appreciation

factors of online community sites. We would thus further a research informed design of community sites, and possibly help members to reach their goals as well as community founders to improve the appreciation of their sites.

ACKNOWLEDGEMENT

We like to thank Ad Feelders for his support in performing latent class analysis.

REFERENCES

Bakker A. (2006) *Kom Verder! Examenboek kennis van de Nederlandse samenleving*. (Come in! Knowledge of the Dutch Society for new habituants) Boom, Amsterdam

Hoevers, J., Meulendijk, M. (2006) *Binary Relations*. University of Utrecht: unpublished b.a. research report

Hummel J, Lechner U (2002) Social Profiles of Virtual Communities. *Proceedings of the 35th Hawaii International Confernce on Systems Sciences*. IEEE

Kim A (2000) *Community Building on the Web*. California: Peachpit Press, Berkely

Kollock P (1997). Design Principles for Online Communities. The Internet and Society: *Harvard Conference Proceedings*, MA: O'Reilly.

Lazarsfeld, P.F., Henry, N.W. (1968) *Latent structure analysis*. Boston: Houghton Mill

Leimeister JM, Sidiras P, Krcmar H (2004) Success factors of virtual communities from the perspective of members and operators: An empirical study. *Proceedings of the 37th Hawaii International Conference on System Sciences*

Preece J, Maloney-Krichmar (2003) Online Communities. J. Jacko and A. Sears (eds.) *Handbook of Human-Computer Interaction*. pp. 596-620: Mahwah,NJ: Lawrence-Erlbaum Associates Inc,

Ridings C, Gefen D (2004) Virtual Community Attraction: Why People Hang Out Online. *Journal of Computer-Mediated Communication* 10 (1)

Ostrom E (1990) Governing the Commons; The Evolution of Institutions for Collective Action. J. E. Alt & D. C. North (eds.) *The political economy of institutions and decisions*. Cambridge University Press, Cambridge.

Oldenburg R (1991) *The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community*. Paragon House, New York

Porter CE (2004) A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. *Journal of Computer-Mediated Communication*, November 2004, 10 (1)

Thij, E. ten (2007) Online communities: Exploring classification approaches using participants' perspectives. In P. Kommers, P. Isaias (Eds.), *Proceedings of the IADIS International Conference on Web Based Communities*

Thij, E. ten, & Wijngaert, L. van de (2006). Using website features to predict Community Website Evaluation. In P. Vicente & Guerrero-Bote (Eds.), *Current Research in Information Sciences and Technologies* (pp. 172-176). Merida: Open Institute of Knowledge.

Thij, E. ten, & Wijngaert, L. van de (2007) Validation of success factors for dance community sites: towards a model for predicting appreciation of online community websites. To appear in: *International Journal of Web Based Communities*

Vermunt, J.K. (1997) *LEM: A General Program for the Analysis of Categorical Data*. Department of Methodology and Statistics, Tilburg University

Van Wendel de Joode, R. (2005). *Understanding open source communities. An organizational perspective.* Ph.d thesis, Delft University of Technology, the Netherlands

Walker CC, Dooley K.J (1999). The Stability Of Self-Organized Rule-Following Work Teams. *Conceptual & Mathematical Organization Theory*, 5 (1).

AGE DECAYING H-INDEX FOR SOCIAL NETWORKS OF CITATIONS

Dimitrios Katsaros[1] Antonis Sidiropoulos[2] Yiannis Manolopoulos[2]
Dept. of Computer & Communication Engineering[1] *University of Thessaly* *Volos, Greece*
[2]*Department of Informatics* *Aristotle University* *Thessaloniki, Greece*

ABSTRACT

Mining the link structure of social networks is a relatively old topic dating back to 1950, but the field has recently gained tremendous attention, due to the compilation of huge amounts of data and the development of sophisticated computerized methods to process these data. Among the many facets of these mining tasks in various types of social networks, the mining task of ranking researchers or conferences/journals in citation networks is of particular importance and challenge due to its uses in many practical situations, e.g., faculty promotion, funding. The development of a fair metric to perform the ranking is a very hard task and a lot of methods have appeared in the literature. Recently, the method that has stimulated the interest of the research community is the h-index method. The h-index metric corrects a lot of the older methods inefficiencies. But, the h-index presents a major disadvantage: it keeps increasing even if the researcher has stopped publishing good papers or if his present work is not that significant. In this paper, we present a generalization of the basic h-index, termed age decaying h-index to cope with this situation. The new index is able to identify the researchers and publication forums that are perceived as very significant for the latest years. We present an extensive experimental evaluation of the proposed novel citation index over the DBLP bibliographic database. The evaluation proves the robustness and virtues of the new index and reveals some very interesting results concerning the performance of scientists, the emergence of research topics, and the preference of scientists to publish in new forums.

KEYWORDS

Hirsch index, h-index, ranking, bibliometrics, link mining, social networks.

1. INTRODUCTION

The academic community works on the evaluation of the scientists since the decade of 1970 (Garfield, 1972). Lately, due to the huge development of the digital libraries, the research efforts on the topic have multiplied, and a large number of researchers is involved in this topic (Barnes, 2005; Bernstein et al., 2005; Bharati and Tarasewich, 2002; Katerattanakul et al., 2003; Kelly Rainer and Miller, 2005; Lowry et al., 2004; Mylonopoulos and Theoharakis, 2001; Nascimento et al., 2003; Nerur et al., 2005; Rahm and Thor, 2005; Schwartz and Russo, 2004; Sidiropoulos and Manolopoulos, 2005*a*).

So far there are two major methods for the scientists' evaluation. The first method is based on the polling. A group of people has to be interviewed for their evaluation. The bigger the sample of people is, the better the evaluation that will be returned is. These works are very interesting, because they perform a ranking according to readers' (and authors') perception, but they suffer from the fact of being basically "manual" and usually biased, and not highly computerized and objective. The second method is based on the social network theory and is conducted through the citation analysis. The evaluation of the scientific work is performed by defining an objective function that calculates some "score" for the "objects" under evaluation, analyzing the social network formed by the citations among the published articles. Defining a quality and representative metric is not an easy task, since it should account for the productivity of a scientist and the impact of all of his/her work (analogously for journals/conferences). Most of the existing methods up-to-date are based on some form of (arithmetics upon) the total number of authored papers, the average number of authored papers per year, the total number of citations, the average number of citations per paper, the average number of citations per year, etc.

Finally, characteristic works implementing the hybrid approach of combining the experts' judge and citation analysis are described in (Kelly Rainer and Miller, 2005; Sidiropoulos and Manolopoulos, 2006). Their rankings are realized by taking some averages upon the results obtained from the citation analysis and experts' opinion, thus implementing a post-processing step of the two major approaches.

1.1 H-index and variations

Although, there is no clear winner among citation analysis and experts' assessment, the former is usually the preferred method, because it can be performed in a fully automated and computerized manner and it is able to exploit the wealth of citation information available in digital libraries.

All the metrics used so far in citation analysis present one or more drawbacks. These drawbacks have been presented by Hirsch (2005) and Sidiropoulos et al. (2007).

To collectively overcome all these disadvantages of the present metrics, during 2005 J. E. Hirsch proposed the pioneering *h-index* (Ball, 2005; Hirsch, 2005), defined as follows[1]:

Definition 1 *A researcher has h-index h if h of his/her N_p articles have received at least h citations each, and the rest $(N_p - h)$ articles have received no more than h citations.*

This metric calculates how broad the research work of a scientist is. The *h-index* accounts for both productivity and impact. For some researcher, to have large *h-index*, s/he must have a lot of "good" articles.

The *h-index* acts as a lower bound on the real number of citations for a scientist. Think that the quantity h will always be smaller than or equal to the number N_p of the articles of a researcher; it holds that $h^2 \leq N_{c,tot}$, where $N_{c,tot}$ is the total number of citations that the researcher has received. Apparently, the equality holds when all the articles, which contribute to *h-index* have received exactly h citations each, which is quite improbable. Therefore, in the

[1]Notice that the economics literature defines the H-index (the Herfindahl-Hirschman index), which is a way of measuring the concentration of market share held by particular suppliers in a market. The H index is the sum of squares of the percentages of the market shares held by the firms in a market. If there is a monopoly, i.e., one firm with all sales, the H index is 10000. If there is perfect competition, with an infinite number of firms with near-zero market share each, the H index is approximately zero. Other industry structures will have H indices between zero and 10000.

usual case it will hold that $h^2 < N_{c,tot}$. To bridge this gap, J. E. Hirsch defined the index a as follows:

Definition 2 *A scientist has a-index a if the following equation holds (Hirsch, 2005):*

$$N_{c,tot} = ah^2. \tag{1}$$

The *a-index* can be used as a second metric-index for the ranking of scientists. It describes the "magnitude" of each scientist's "hits". A large a implies that some article(s) have received a fairly large number of citations compared to the rest of its articles.

The introduction of the *h-index* was a major breakthrough in citation analysis. Though several aspects of the inefficiency of the original *h-index* are apparent; or to state it in its real dimension, significant efforts are needed to unfold the full potential of *h-index*. Firstly, the original *h-index* assigns the same importance to all citations, no matter what their age is, thus refraining from revealing the trendsetters scientists. Secondly, the *h-index* assigns the same importance to all articles, thus making the young researchers to have a relatively small *h-index*, because they did not have enough time either to publish a lot of good articles, or time to accumulate large number of citations. Thus, the *h-index* can not reveal the brilliant though young scientists.

After the introduction of the *h-index*, a number of other proposals followed, either presenting case studies using it (Bar-Ilan, 2006; Braun et al., 2005; Rousseau, 2006), or describing a new variation of it (Egghe, 2006*b*) (aiming to bridge the gap between the lower bound of total number of citations calculated by *h-index* and their real number), or studying its mathematics and its performance (Bornmann and Daniel, 2005; Egghe, 2006*a*). The interested reader can find a survey of the articles about *h-index* in Bornmann and Daniel (2007).

Deviating from their line of research, Sidiropoulos et al. (2007) developed a pair of generalizations of the *h-index* for ranking scientists, which are novel citation indices, a normalized variant of the *h-index* and a pair of variants of the *h-index* suitable for journal/conference ranking.

1.1.1 The contemporary h-index

The original *h-index* does not take into account the "age" of an article. It may be the case that some scientist contributed a number of significant articles that produced a large *h-index*, but now s/he is rather inactive or retired. Therefore, senior scientists, who keep contributing nowadays, or brilliant young scientists, who are expected to contribute a large number of significant works in the near future but now they have only a small number of important articles due to the time constraint, are not distinguished by the original *h-index*. Thus, arises the need to define a generalization of the *h-index*, in order to account for these facts.

We have defined a score $S_c(i)$ for an article i based on citation counting, as follows:

$$S_c(i) = \gamma * (Y(now) - Y(i) + 1)^{-\delta} * |C(i)| \tag{2}$$

where $Y(i)$ is the publication year of article i and $C(i)$ are the articles citing the article i. If we set δ=1, then $S_c(i)$ is the number of citations that the article i has received, divided by the "age" of the article. Since, we divide the number of citations with the time interval, the quantities $S_c(i)$ will be too small to create a meaningful *h-index*; thus, we use the coefficient γ. In the experiments reported by Sidiropoulos et al. (2007) the value of 4 is used for the coefficient γ and

the value of 1 for δ. In Section 3. we will use the same values. Thus, for an article published during the current year, its citations account four times. For an article published 4 year ago, its citations account only one time. For an article published 6 year ago, its citations account $\frac{4}{6}$ times, and so on.

This way, an old article gradually loses its "value", even if it still gets citations. In other words, in the calculations we mainly take into account the newer articles[2]. Therefore, we define a novel citation index for scientist rankings, the *contemporary h-index*, expressed as follows:

Definition 3 *A researcher has contemporary h-index h_c, if h_c of its N_p articles get a score of $S_c(i) \geq h_c$ each, and the rest $(N_p - h_c)$ articles get a score of $S_c(i) \leq h_c$.*

1.1.2 The trend h-index

The original *h-index* does not take into account the year when an article acquired a particular citation, i.e., the "age" of each citation. For instance, consider a researcher who contributed to the research community a number of really brilliant articles during the decade of 1960, which, say, got a lot of citations. This researcher will have a large *h-index* due to the works done in the past. If these articles are not cited anymore, it is an indication of an outdated topic or an outdated solution to the problem. On the other hand, if these articles continue to be cited, then we have the case of an *influential mind*, whose contributions continue to shape newer scientists' minds. There is also a second very important aspect in aging the citations. There is the potential of disclosing *trendsetters*, i.e., scientists whose work is considered pioneering and sets out a new line of research that currently is hot ("trendy"), thus this scientists' works are cited very frequently.

To handle this, we take the opposite approach than *contemporary h-index*'s; instead of assigning to each scientist's article a decaying weight depending on its age, we assign to each citation of an article an exponentially decaying weight, which is as a function of the "age" of the citation. This way, we aim at estimating the impact of a researcher's work in a particular time instance. We are not interested in how old the articles of a researcher are, but whether they still get citations. We define an equation similar to Equation 2, which is expressed as follows:

$$S_t(i) = \gamma * \sum_{\forall x \in C(i)} (Y(now) - Y(x) + 1)^{-\delta} \tag{3}$$

where γ, δ, $Y(i)$ and $S(i)$ for an article i are as defined earlier. We define a novel citation index for scientist ranking, the *trend h-index*, expressed as follows:

Definition 4 *A researcher has trend h-index h_t if h_t of its N_p articles get a score of $S_t(i) \geq h_t$ each, and the rest $(N_p - h_t)$ articles get a score of $S_t(i) \leq h_t$ each.*

Apparently, for $\gamma = 1$ and $\delta = 0$, the *trend h-index* coincides with the original *h-index*.

1.2 Our contributions

The purpose of our work is to extend and generalize the original *h-index* and its variations in such ways, so as to reveal various latent though strong facts hidden in citation networks. In this context, the article makes the following contributions:

[2] Apparently, if δ is close to zero, then the impact of the time penalty is reduced, and, for $\delta = 0$, this variant coincides with the original *h-index* for $\gamma = 1$.

- Introduces a generalization of the *h-index*, namely the *age decaying h-index*, which is appropriate for scientist ranking and is able to reveal *brilliant young scientists* and *trend-setters*. This generalization can also be used for conferences and journals ranking.
- Performs an extensive experimental evaluation of the aforementioned citation indices, using real data from DBLP, an online bibliographic database.

The rest of this article is organized as follows: In Section 2., we present the novel citation index *age decaying h-index*, and in Section 3. presents the evaluation of the introduced citation index against its predecessors. Finally, Section 4. summarizes the paper's contributions and concludes the article.

2. A NOVEL CITATION INDEX FOR SCIENTIST, CONFERENCES AND JOURNALS RANKING

2.1 The age decaying h-index

The *trend h-index* takes into account the "age" of the citations. On the on the hand *contemporary h-index* takes into account the "age" of the publications. The *age decaying h-index* is a generalization of both the *contemporary h-index* and *trend h-index*, which takes into account both the age of a scientist's article and the age of each citation to his/her articles.

We define a score function S_{ad} for a publication i as:

$$S_{ad}(i) = \gamma^2 * (Y(now) - Y(i) + 1)^{-\delta_1} * \sum_{\forall x \in C(i)} (Y(now) - Y(x) + 1)^{-\delta_2} \quad (4)$$

where γ, δ_1, δ_2 and $Y(i)$ for an article i are as defined earlier. If δ_1 and δ_2 are equal, then the "age" of the publication and the "age" of the citation have the same importance. We may give greater importance to one of them by increasing the corresponding δ (δ_1 or δ_2).

We define a novel citation index for scientist ranking, the *age decaying h-index*, expressed as follows:

Definition 5 *A researcher has age decaying h-index h_{ad} if h_{ad} of its N_p articles get a score of $S_{ad}(i) \geq h_{ad}$ each, and the rest $(N_p - h_{ad})$ articles get a score of $S_{ad}(i) \leq h_{ad}$ each.*

Likewise, the *age decaying h-index* can be defined for a Journal or a Conference. For instance, the *age decaying h-index* of a journal/magazine or a Conference is h_{ad}, if h_{ad} of the N_p articles that contains, have received at least h_{ad} citations each, and the rest $(N_p - h_{ad})$ articles received no more than h_{ad}.

The second metric of the original *h-index* notion is the factor a. Factor a_{ad} can be defined as:

$$\sum_{\forall i \in P} S_t(i) = a_{ad} * h_{ad}^2 \quad (5)$$

where P is the set of a scientist's publications. The a-index can be used as a second metric-index for the evaluation and ranking of scientists. It describes the age decaying "magnitude" of each scientist's "hits". A large a implies that some article(s) have received a fairly large number of citations compared to the rest of its articles and with respect to what the h-index presents.

3. EXPERIMENTS

Having defined this generalization and variants of the original *h-index*, we will evaluate in the subsequent sections their success in identifying scientists or forums with extraordinary performance or their ability to reveal latent facts in a citation network, such as brilliant young scientists and trendsetters. For the evaluation, we will exploit the on-line database of DBLP [3].

In the sequel, we will present a small subset of the results obtained for ranking scientists, conferences and journals, using the basic *h-index* definition as well as using the generalization developed in the previous section. Along the lines of (Sidiropoulos and Manolopoulos, 2005*a*,*b*, 2006), our dataset consists of the DBLP collection (DBLP timestamp: Mar/3/2006). The reason for selecting this source of data instead of ISI or Google data is twofold:

1. DBLP contains data about journal and conference publications as well, and
2. DBLP data are focused mostly in the area of Databases.

It is worthwhile noticing that many top conferences of this area are very competitive (with an acceptance ratio stronger than 1:3 and up to 1:7), and occasionally more competitive that the top journals of the area. In many computer science departments worldwide, publications in these conferences are favored in comparison to journal publications. Therefore, a ranking of conferences on databases is equally important to the ranking of the journals of the area.

The reason for selecting this "old" snapshot of the DBLP database is to be able to compare the results with our former published research. The used database snapshot contains 451694 inproceedings, 266307 articles, 456511 authors, 2024 conference series and 504 journals. Also, the number of citations in our dataset is 100205. Although this number is relatively small, it is a satisfactory sample for our purposes. Almost all citations in the database are made from publications prior to the year 2001. Thus, we can assume that the results presented here correspond to the year 2001. From now on, with the term "now" we actually mean sometime near 2001. Although other bibliographic sources, like ISI, are widely available and much more complete, the used collection has the two above desired characteristics and thus it is sufficient for exhibiting the benefits of our proposed citation indices, without biasing our results.

3.1 Experiments with the h-index for scientists

In Tables 1 and 2 we present the resulting ranking using the *h-index*, as well as its defined generalization, the *age decaying h-index*. In these tables column a_{ad} stands for the factor a of the *age decaying h-index*. Table 1 is sorted by the *h-index* ranking position. In this table we also present the values for *contemporary h-index* (h_c), *trend h-index* (h_t) and *age decaying h-index* (h_{ad}) and the corresponding rank position (sub-columns @ pos). For example, at the first position is ranked Michael Stonebraker with *h-index* 24, $a = 3.78$, total number of citations equal to 2180, total number of published papers =193, *age decaying h-index* equals 11 and his corresponding position at the *age decaying h-index* rank table is position number 14, *contemporary h-index* equals 13 and his position with the *contemporary h-index* metric is number 3, . . .

At a first glance, we see that the values computed for *h-index* (Table 1) are much lower than the values presented in (Hirsch, 2005) for physics scientists due to the non completeness of the

[3] The DBLP digital library with bibliographic data on "Databases and Logic Programming" is maintained by Michael Ley at the University of Trier, accessible from `http://dblp.uni-trier.de/`

Table 1. Scientist ranking with *h-index*.

Name	h	a	$N_{c,tot}$	N_p	h_{ad}(@ pos)	h_c(@ pos)	h_t(@ pos)
1.Michael Stonebraker	24	3.78	2180	193	11(@ 14)	13(@ 3)	19(@ 3)
2.Jeffrey D. Ullman	23	3.37	1783	227	14(@ 6)	14(@ 2)	20(@ 2)
3.David J. DeWitt	22	3.91	1896	150	14(@ 7)	16(@ 1)	23(@ 1)
4.Philip A. Bernstein	20	3.39	1359	124	7(@ 73)	10(@ 15)	12(@ 23)
5.Won Kim	19	2.96	1071	143	7(@ 71)	10(@ 12)	14(@ 12)
6.Catriel Beeri	18	3.16	1024	93	7(@ 66)	10(@ 13)	13(@ 18)
7.Rakesh Agrawal	18	3.06	994	154	16(@ 1)	13(@ 4)	19(@ 4)
8.Umeshwar Dayal	18	2.81	913	130	8(@ 45)	9(@ 20)	13(@ 16)
9.Hector Garcia-Molina	17	3.60	1041	314	13(@ 9)	10(@ 8)	17(@ 7)
10.Yehoshua Sagiv	17	3.52	1020	121	9(@ 35)	9(@ 18)	13(@ 14)
11.Ronald Fagin	17	2.83	818	121	5(@ 130)	7(@ 48)	11(@ 38)
12.Jim Gray	16	6.13	1571	118	11(@ 16)	11(@ 7)	14(@ 10)
13.Serge Abiteboul	16	4.33	1111	172	16(@ 3)	12(@ 5)	17(@ 6)
14.Michael J. Carey	16	4.25	1090	151	10(@ 22)	10(@ 9)	14(@ 11)
15.Nathan Goodman	16	3.37	865	68	5(@ 161)	7(@ 49)	10(@ 49)
16.Christos Faloutsos	16	2.89	742	175	13(@ 10)	10(@ 11)	17(@ 8)
17.Raymond A. Lorie	15	6.23	1403	35	5(@ 134)	8(@ 29)	11(@ 33)
18.Jeffrey F. Naughton	15	2.90	653	123	14(@ 8)	10(@ 10)	15(@ 9)
19.Bruce G. Lindsay	15	2.76	623	60	6(@ 91)	8(@ 37)	12(@ 32)
20.David Maier	14	5.56	1090	158	8(@ 49)	10(@ 14)	12(@ 24)

Table 2. Scientist ranking with *age decaying h-index*.

Name	h_{ad}	a_{ad}	$N_{c,tot}$	N_p	h(@ pos)	h_c(@ pos)	h_t(@ pos)
1.Rakesh Agrawal	16	3.28	994	154	18(@ 7)	13(@ 4)	19(@ 4)
2.Jennifer Widom	16	3.19	709	136	14(@ 23)	12(@ 6)	18(@ 5)
3.Serge Abiteboul	16	3.08	1111	172	16(@ 13)	12(@ 5)	17(@ 6)
4.Dan Suciu	16	2.79	244	113	9(@ 100)	9(@ 22)	12(@ 25)
5.Alon Y. Levy	15	2.85	321	77	10(@ 69)	9(@ 21)	14(@ 13)
6.Jeffrey D. Ullman	14	4.18	1783	227	23(@ 2)	14(@ 2)	20(@ 2)
7.David J. DeWitt	14	3.41	1896	150	22(@ 3)	16(@ 1)	23(@ 1)
8.Jeffrey F. Naughton	14	2.95	653	123	15(@ 18)	10(@ 10)	15(@ 9)
9.Hector Garcia-Molina	13	4.07	1041	314	17(@ 9)	10(@ 8)	17(@ 7)
10.Christos Faloutsos	13	2.62	742	175	16(@ 16)	10(@ 11)	17(@ 8)
11.Daniela Florescu	13	2.44	105	60	5(@ 324)	8(@ 43)	9(@ 69)
12.Hans-Peter Kriegel	12	3.26	465	204	11(@ 50)	8(@ 28)	12(@ 21)
13.Joseph M. Hellerstein	12	2.76	252	86	10(@ 79)	8(@ 36)	12(@ 31)
14.Michael Stonebraker	11	4.12	2180	193	24(@ 1)	13(@ 3)	19(@ 3)
15.H. V. Jagadish	11	3.59	503	151	12(@ 39)	10(@ 16)	13(@ 17)
16.Jim Gray	11	3.58	1571	118	16(@ 12)	11(@ 7)	14(@ 10)
17.Surajit Chaudhuri	11	3.22	263	114	9(@ 97)	8(@ 34)	12(@ 30)
18.Yannis Papakonstantinou	11	3.06	219	57	8(@ 124)	8(@ 39)	10(@ 48)
19.Tova Milo	11	2.53	179	74	8(@ 133)	8(@ 41)	9(@ 64)
20.Leonid Libkin	11	2.46	143	99	6(@ 248)	6(@ 78)	10(@ 52)

source data. Also, we can notice that the values for h and h_{ad} are different with each other as well as there are differences in the ordering of the scientists. This confirms our allegation for the convenience of these indices.

In contrast with our contemporary and trend h-index research (Sidiropoulos et al., 2007), Tables 1 and 2 present significant differences. The rank order of Table 1 is expected, since well known names of the database domain are ranked at the first 20 positions. On the other hand, Table 2 presents a different ordering with new names appeared at the first 20 positions. The researchers that are reported to be in the top 20 with the *age decaying h-index* but not with the original *h-index* are: Dan Suciu, Alon Y. Levy, Daniela Florescu, Hans-Peter Kriegel, Joseph Hellerstein, H. V. Jagadish, Surajit Chaudhuri, Yannis Papakonstantinou, Tova Milo and Leonid Libkin. Also, the ordering given by *age decaying h-index* is different than the ones of *contemporary h-index* and *trend h-index*. This fact confirms that the *age decaying h-index* is a

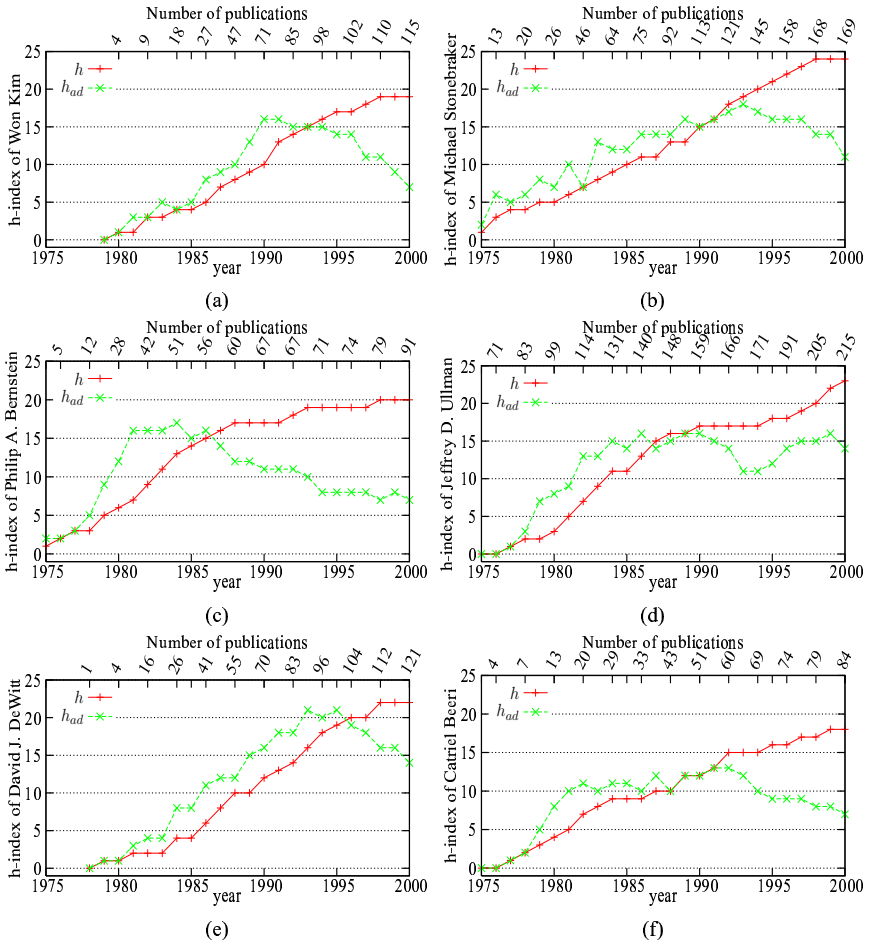

Figure 1. The *h-index* of scientists working in databases area.

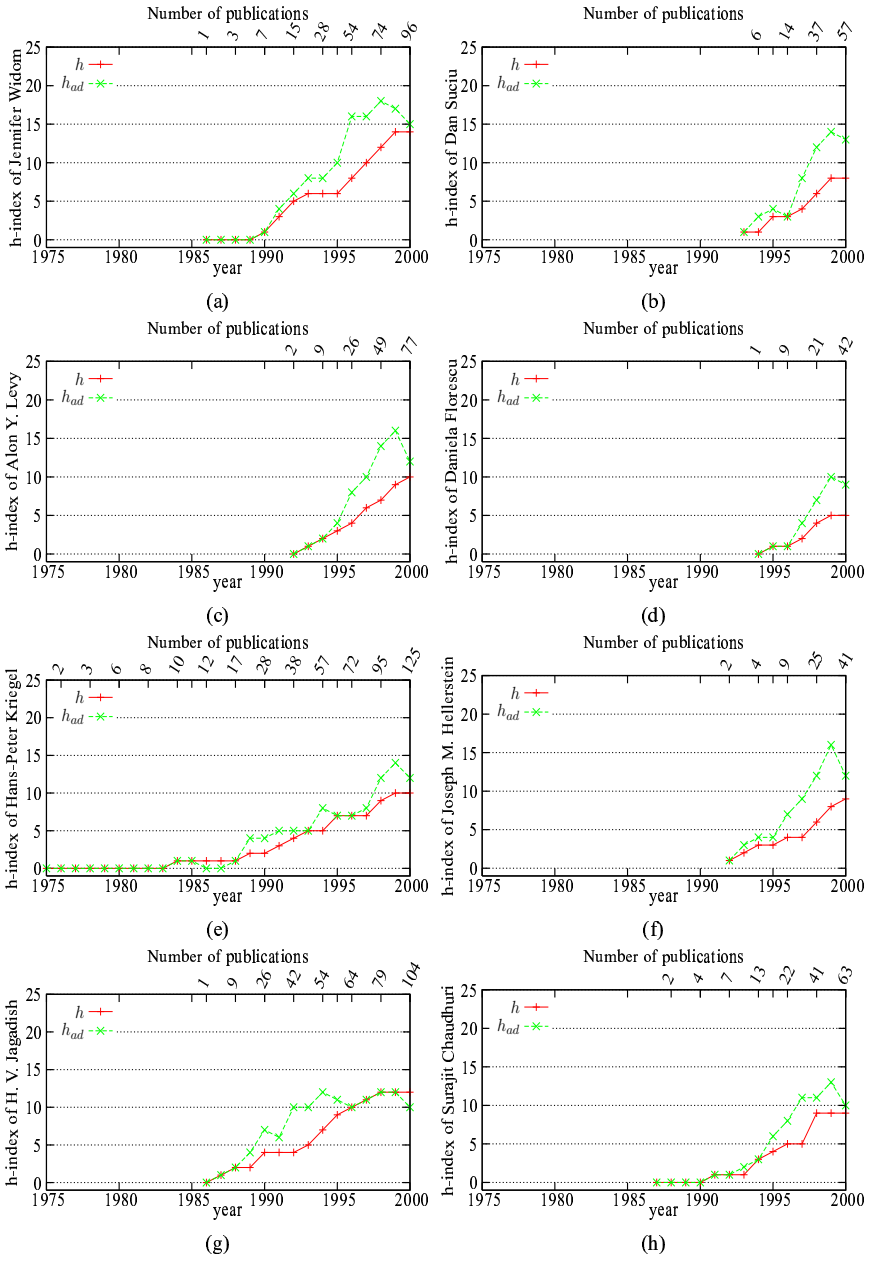

Figure 2. The *h-index* of scientists working in databases area (part b).

novel method. The majority of the "new" scientists at the top 20 positions, can be said that are "young" scientists compared to the "all time classics" scientists. This can also be confirmed from the Figures 2 and 1. As we can see in Figure 2 most of the "new" ones have started published

around 1990, in contrast with the scientists presented in Figure 1 who they started publishing around 1975. This means that our new index is really age decaying. Thus, it assists the scientists with new publications and simultaneously new citations.

It is also worthwhile to mention that the *contemporary h-index* and *trend h-index* are fair metrics for the "all-time classic" scientists, e.g., Jeffrey Ullman, Michael Stonebraker, and David DeWitt, whose influential works continue to shape the modern scientists way of thinking.

Motivated by the differences in the above tables, we present the collection of graphs in Figure 1. In these figures, we can see the history of the *h-index* for those scientists, who present significant differences between the *h-index* family of citation indices, and also those who have a rapid upward slope at their plot curves. Again, we remind that our data set is rather incomplete for the years after 2000, and thus a downwards pitch for all the researchers appears during the years 1999-2000. However, the results are indicative.

Won Kim (Figure 1(a)), Michael Stonebraker (Figure 1(b)) and Philip A. Bernstein (Figure 1(c)) present a similar path. For instance, there is a high ascending curve for *age decaying h-index* until around 1990 (with few years difference). Therefore, we expect that *h-index* will not present high increase. This is explained by the fact that the main research interests of Won Kim was on object-oriented database systems, which flourished during the last years of the eighties and in the first years of the nineties, but later become a relatively inactive area. Stonebraker and Bernstein, after their intensive and high quality research, which laid the foundations of the relational model during the '80s, reduced their productivity.

Jeffrey D. Ullman's h_{ad} followed an uprising course until 1985, then started to be stabilized and lightly decreasing, but after 1994 it started increasing again. This is due to the fact that at that time, J. D. Ullman worked with his colleagues on the integration of distributed data sources and particularly his research focused on semistructured data, that happened to be very popular and trendy research theme.

The pattern of increase of the age decaying h-index for David DeWitt and Catriel Beeri is quite similar, with a shift of a few years in the time scale, both of whom, after fundamental contributions to the theory and practice of the relational model that brought them at the forefront of the research, did not dealt with the new research topics that emerged at that time.

In Figure 2(a), we see the progress rate for Jennifer Widom. While Jennifer Widom is not even among the top 20 researchers using the *h-index*, she is on the 2^{nd} position using the *age decaying h-index*. Also, she is ranked 6^{th} and 5^{th} using the *contemporary h-index* and *trend h-index*, respectively. She is one of the researchers from our list that presents such a big difference on the timing rate compared to the basic *h-index*. As we can also see from the diagram, this difference is justifiable, since the increase rate of the basic *h-index* is high. Jenifer Widom made some ground breaking contributions on building semistructured data management systems, that laid the foundations for the modern XML management systems.

Dan Suciu climbed from the 100^{th} place by the original *h-index* to the 4^{th} by the *age decaying h-index*. Figure 2(b) shows that the *age decaying h-index* follows a rapidly ascending course, as well as that for Alon Y. Levy presented in Figure 2(c). Daniela Florescu gained the highest rise from all the scientists presented in this paper. She is ranked at the 324^{th} place by the original *h-index* and moved to the 11^{th} position. The pattern of growth of all these scientists is not accidental; all of them have worked on the topic of semistructured data, which later was transformed to the area of XML data management, which can be easily recognized as one of the most hot and trendy topics during the last years of the previous decade and the first years of this

decade.

Joseph M. Hellerstein (Figure 2(f)) and Surajit Chaudhuri (Figure 2(h)) follow a similar slope. Although both researchers have broad research interests, it is easy to ascribe the growth of their *age decaying h-index* to their contributions to the relational databases and to online analytical processing (OLAP) and data warehousing.

Hans-Peter Kriegel (see Figure 2(e)) has been recognized as one of the most productive researchers in the area of spatial data management; this topic was very popular and attracted a lot of interest during the previous decade. Therefore, the pattern of growth of his *age decaying h-index* is reasonable. Similarly, the *age decaying h-index* of H. V. Jagadish, who was working at that time on multidimensional data, exhibits similar growth pattern.

Collectively, starting from the observations about the scientists with steep growing of their *age decaying h-index*, we can go one step further and recognize research topics which constitute the preferred and trendy research areas at that periods, like spatial data, semistructured data and OLAP. Indeed, the findings of our citations indexes are in absolute accordance with what the common sense deduces by observing the number of paper on each topic in major journals and conferences. Thus, the proposed citation index is able to reveal large scientific areas as promising topics for research.

3.2 Experiments with conferences and journals ranking

3.2.1 Experiments with conferences ranking

To evaluate our citation indices on conference ranking, we extract only the database conferences (as defined by Elmacioglu and Lee (2005)) from the data we used in the previous section. In this section we will make experiments using the indicator that we fixed for scientists, namely *h-index* and *age decaying h-index*.

In Table 3 we present the top-10 conferences using the *h-index* for the ordering. The symbol a in Table 3 and the symbol a_{ad} in Table 4 correspond to the *a-index* on Definition 2 and Equation 5 respectively. Since the quality of the conferences is relatively constant, we observe that in Tables 3 and 4 there are no significant differences in the ranking. The differences occur below the 5^{th} place where "International Conference on Conceptual Modeling (ER)" and "Expert Database Systems (EDS)" are replaced by "International Conference on Database Theory (ICDT)" and "Knowledge Discovery and Data Mining (KDD)".

In Figure 3 we present in the same way we used for scientists, the progress of selected conferences. Note here that the *h-index* is shown per year in the graphs, which means that this is the computed *h-index* during the specific year. E.g., the *h-index* that is computed for the VLDB for 1995 is the *h-index* that is computed if we exclude everything from our database after 1995.

Due to the lack of citations for the years after 1999, in all graphs there is a stabilization of the *h-index* line and a downfall for the indicator *age decaying h-index*. Figure 3(a) presents the history of the SIGMOD conference. According to the tables, SIGMOD is ranked first. In the figure, we observe its steeply ascending line as well as the *age decaying h-index* remains higher than the *h-index* (until 1999). Also, VLDB (Figure 3(b)) follows an ascending path. These two conferences are clearly ranked first by our algorithm and by *h-index*. On the other hand, the PODS conference (Figure 3(c)) follows a bending line after 1988 with some picks. ICDE is a relatively younger conference compared to the rest of the conferences presented, but we can see in the plot (Figure 3(d)), that it follows a rapidly ascending course until 1987 and afterwords it's

age decaying h-index is almost stabilized with an increasing trend.

Finally, with respect to the ADBT conference (Figure 3(e)) we mention that this conference was organized only three times (1977, 1979 and 1982). As we can see in the upper x axis, the number of publications stops increasing after 1982. Thus, we can not compare it to the rest of the conferences. What we observe from this plot, is that the *age decaying h-index* converges to zero which confirms the correctness of our metric.

KDD is the "youngest" conference among the rest, but it has managed to climb up to the 6^{th} place in the *age decaying h-index* rank table. From the plot (Figure 3(f)) we cannot gather much more information due to its short history and the lack of available data.

3.2.2 Experiments with journals ranking

In the case of journals, we can use the basic form of *h-index* as well as the generalization *age decaying h-index* we defined for scientists and for conferences.

Tables 5 and 6 present the top-10 journals according to the aforementioned indices. As expected, the ACM TODS (tods), IEEE TKDE (tkde), SIGMOD Record (sigmod) are the top three journals. The striking observation is that the Information Systems (is) drops in the ranking with the *age decaying h-index*, as compared to its position with *h-index*, implying that it is not considered a prestigious journal anymore; it is ranked even below the Data Engineering Bulletin!,

Table 3. Conferences ranking with *h-index*.

Name	h	a	$N_{c,tot}$	N_p
1.sigmod	45	6.05	12261	2059
2.vldb	37	7.10	9729	2192
3.pods	26	5.74	3883	776
4.icde	22	6.83	3307	1970
5.er	16	5.80	1486	1338
6.edbt	13	3.89	658	434
7.eds	12	3.65	527	101
8.adbt	12	2.86	412	42
9.icdt	11	4.79	580	313
10.oodbs	11	3.96	480	122

Table 4. Conferences ranking with *age decaying h-index*.

Name	h_{ad}	a_{ad}	$N_{c,tot}$	N_p	h
1.sigmod	32	5.85	12261	2059	45
2.vldb	25	6.94	9729	2192	37
3.pods	20	5.32	3883	776	26
4.icde	17	8.01	3307	1970	22
5.icdt	12	5.27	580	313	11
6.kdd	11	4.08	243	1074	6
7.edbt	11	3.92	658	434	13
8.webdb	9	2.69	31	163	3
9.cikm	8	4.18	211	1030	7
10.ssdbm	8	3.71	321	609	7

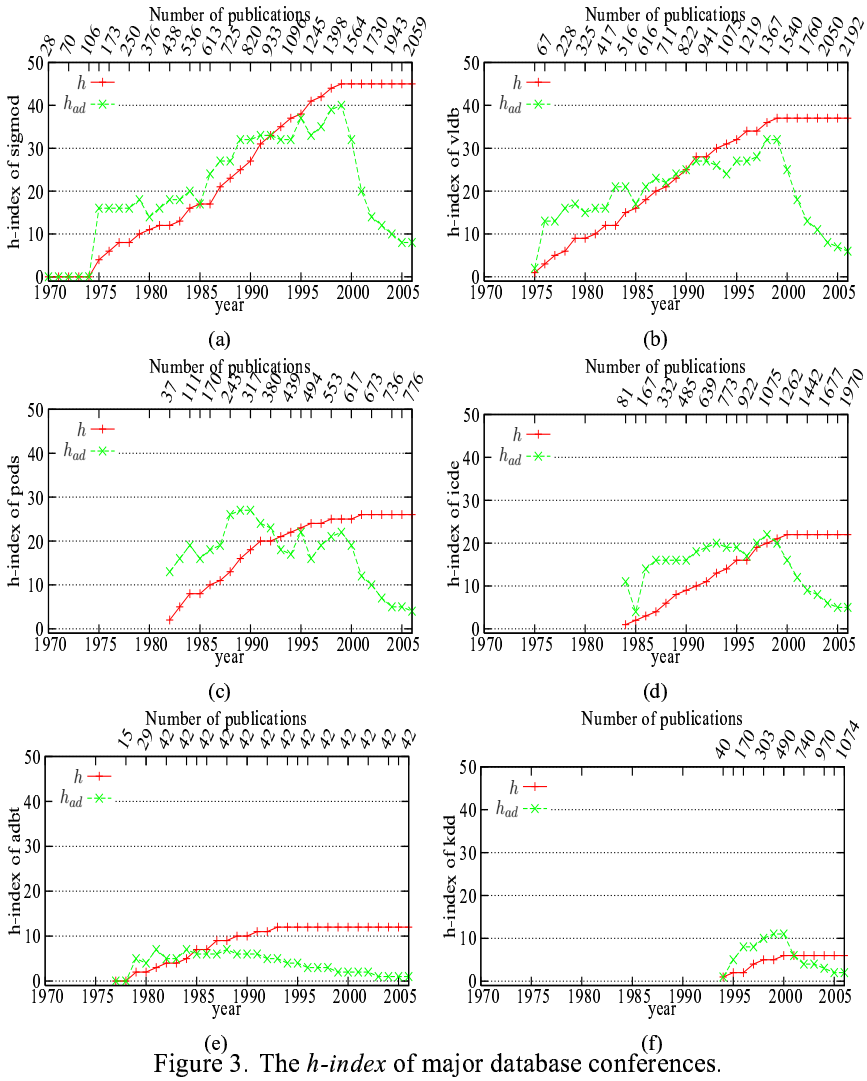

Figure 3. The *h-index* of major database conferences.

implying that it is not publishing modern research results as it used to. On the contrary, SIGMOD Record and VLDB Journal (vldb) show an uprising trend.

In Figure 4 we present the results of computing the defined indices for the major journals of the database domain on a per year basis. Due to the lack of available data after the year 2000, all indices drop steeply. Though, the case of ACM TODS is worthwhile mentioning. Its *age decaying h-index* (Figure 4(a)) drops after 1993, which can be attributed to the relatively large end-to-end publication time of its articles during the years 1990-2000 (Snodgrass, 2003), which acted as an impediment for the authors to submit their works in that venue. Fortunately, this is not the case anymore. On the other hand, SIGMOD Record (Figure 4(c)) and VLDB Journal (Figure 4(d)) show a clear uprising trend until 1998. Also, the case of SIGMOD Record is

characteristic, because, even though it has been published since 1970, its indices get really noticeable only after 1980, when this newsletter started to publish some very good survey-type articles and was freely available on the Web, which improved its visibility. Finally, Information Systems (is: Figure 4(e)) and ACM Transactions on Information Systems (tois: Figure 4(f)) show a stable performance based on the *age decaying h-index* (of course by ignoring the years after 1999 due to the lack of data).

4. CONCLUSIONS

Estimating the significance of a scientist's work is a very important issue for prize awarding, faculty recruiting; similarly, the estimation of a publication forum's (journal or conference) is significant since it impacts the scientists' decisions about where to publish their work. This issue has received some attention during the last years, but the interest on this topics has been renewed by a path-breaking paper by J. E. Hirsch, who proposed the *h-index* to perform fair ranking of scientists, avoiding many of the drawbacks of the earlier bibliographic ranking methods.

The initial proposal and meaning of the *h-index* has various shortcomings, mainly of its inability to differentiate between active and inactive (or retired) scientists and its weakness to

Table 5. Journal ranking with *h-index*.

Name	h	a	$N_{c,tot}$	N_p
1.tods	49	3.88	9329	598
2.tkde	18	4.69	1520	1388
3.is	16	4.71	1208	934
4.sigmod	15	5.07	1142	1349
5.tois	13	4.37	740	378
6.debu	11	7.13	863	877
7.vldb	9	5.03	408	281
8.ipl	8	6.06	388	4939
9.dke	6	8.77	316	773
10.dpd	6	5.25	189	238

Table 6. Journal ranking with *age decaying h-index*.

Name	h_{ad}	a_{ad}	$N_{c,tot}$	N_p	h
1.tods	13	7.71	9329	598	49
2.sigmod	13	4.94	1142	1349	15
3.tkde	12	5.77	1520	1388	18
4.debu	12	3.49	863	877	11
5.vldb	12	2.82	408	281	9
6.dpd	7	3.82	189	238	6
7.is	6	7.51	1208	934	16
8.jiis	6	5.67	156	318	6
9.tois	5	7.14	740	378	13
10.dke	5	6.52	316	773	6

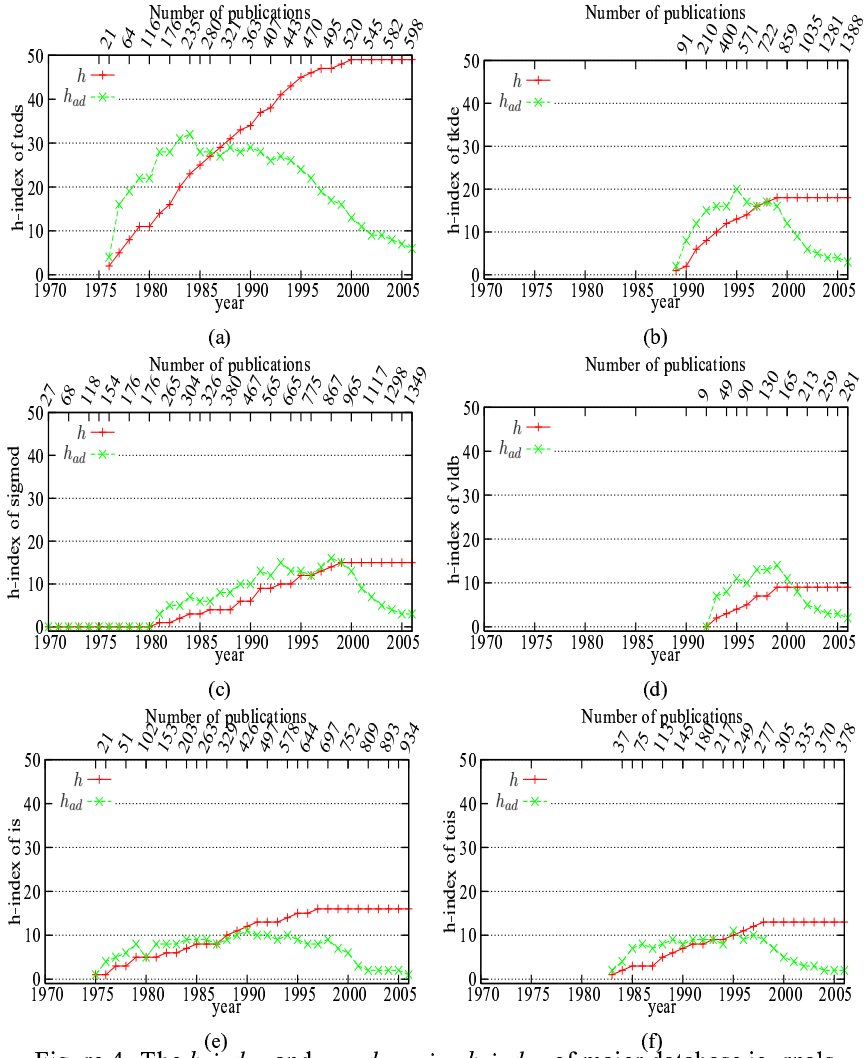

Figure 4. The *h-index* and *age decaying h-index* of major database journals.

differentiate between significant works in the past (but not any more) and the works which continue to shape the scientific thinking.

Based on the identification of these shortcomings of *h-index*, we proposed in this article an effective *age decaying h-index* generalization. This novel citation index aim at the ranking of scientists by taking into account both the age of the published articles as well as the age of the citations to each article.

To evaluate the proposed ranking metrics, we conducted extensive experiments on an online bibliographic database containing data from journal and conference publications as well, and moreover focused in the specific area of databases. From the results we obtained, we concluded that *h-index* is not a general purpose indicative metric. The *age decaying h-index* is able to

disclose latent facts in citation networks, like trendsetters and brilliant young scientists. For the case of conference and journal ranking, the index *age decaying h-index* gives a more fair view for the ranking.

REFERENCES

Ball, P. (2005), "Index aims for fair ranking of scientists – h-index sums up publication record", *Nature* , Vol. 436, p. 900.

Bar-Ilan, J. (2006), "h-index for price medalists revisited", *ISSI Newsletter* , Vol. 5.

Barnes, S. J. (2005), "Assessing the value of IS journals", *Communications of the ACM* , Vol. 48, pp. 110–112.

Bernstein, P. A., Bertino, E., Heuer, A., Jensen, C. J., Meyer, H., Tamer Ozsu, M., Snodgrass, R. T. and Whang, K.-Y. (2005), "An apples-to-apples comparison of two database journals", *ACM SIGMOD Record* , Vol. 34, pp. 61–64.

Bharati, P. and Tarasewich, P. (2002), "Global perceptions of journals publishing e-commerce research", *Communications of the ACM* , Vol. 45, pp. 21–26.

Bornmann, L. and Daniel, H.-D. (2005), "Does the h-index for ranking of scientists really work?", *Scientometrics* , Vol. 65, pp. 391–392.

Bornmann, L. and Daniel, H.-P. (2007), "What do we know about the h-index?", *Journal of the American Society of Information Science and Technology* . to appear.

Braun, T., Glanzel, W. and Schubert, A. (2005), "A Hirsch-type index for journals", *The Scientist* , Vol. 19, pp. 8–10.

Egghe, L. (2006*a*), "Dynamic h-index: The Hirsch index in function of time", *Scientometrics* . to appear.

Egghe, L. (2006*b*), "Theory and practise of the g-index", *Scientometrics* , Vol. 69, pp. 131–152.

Elmacioglu, E. and Lee, D. (2005), "On six degrees of separation in DBLP-DB and more", *ACM SIGMOD Record* , Vol. 34, pp. 33–40.

Garfield, E. (1972), "Citation analysis as a tool in journal evaluation", *Science* , Vol. 178, pp. 471–479.

Hirsch, J. E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences* , Vol. 102, pp. 16569–16572.

Katerattanakul, P., Han, B. T. and Hong, S. (2003), "Objective quality ranking of computing journals", *Communications of the ACM* , Vol. 46, pp. 111–114.

Kelly Rainer, R. and Miller, M. D. (2005), "Examining differences across journal rankings", *Communications of the ACM* , Vol. 48, pp. 91–94.

Lowry, P., Romans, D. and Curtis, A. (2004), "Global journal prestige and supporting disciplines: A scientometric study of information systems journals", *Journal of the Association for Information Systems* , Vol. 5, pp. 29–75.

Mylonopoulos, N. A. and Theoharakis, V. (2001), "Global perception of IS journals", *Communications of the ACM* , Vol. 44, pp. 29–33.

Nascimento, M., Sander, J. and Pound, J. (2003), "Analysis of SIGMOD's co-authorship graph", *ACM SIGMOD Record* , Vol. 32, pp. 8–10.

Nerur, S. P., Sikora, R., Mangalaraj, G. and Balijepally, V. (2005), "Assessing the relative influence of journals in a citation network", *Communications of the ACM* , Vol. 48, pp. 71–74.

Rahm, E. and Thor, A. (2005), "Citation analysis of database publications", *ACM SIGMOD Record* , Vol. 34, pp. 48–53.

Rousseau, R. (2006), "A case study: Evolution of JASIS' Hirsch index", *Library and Information Science* . http://eprints.rcils.org/archive/00005430.

Schwartz, R. B. and Russo, M. C. (2004), "How to quickly find articles in the top IS journals", *Communications of the ACM* , Vol. 47, pp. 98–101.

Sidiropoulos, A., Katsaros, D. and Manolopoulos, Y. (2007), "Generalized Hirsch h-index for disclosing latent facts in citation networks", *Scientometrics* , Vol. 72. to appear.

Sidiropoulos, A. and Manolopoulos, Y. (2005*a*), "A citation-based system to assist prize awarding", *ACM SIGMOD Record* , Vol. 34, pp. 54–60.

Sidiropoulos, A. and Manolopoulos, Y. (2005*b*), "A new perspective to automatically rank scientific conferences using digital libraries", *Information Processing & Management* , Vol. 41, pp. 289–312.

Sidiropoulos, A. and Manolopoulos, Y. (2006), "Generalized comparison of graph-based ranking algorithms for publications and authors", *Journal for Systems and Software* , Vol. 79, pp. 1679–1700.

Snodgrass, R. (2003), "Journal relevance", *ACM SIGMOD Record* , Vol. 32, pp. 11–15.

MAILING LISTS MEET THE SEMANTIC WEB

Sergio Fernández and Diego Berrueta *Fundación CTIC, Parque Científico y Tecnológico, Cabueñes, Gijón, Spain. sergio.fernandez@fundacionctic.org, diego.berrueta@fundacionctic.org*

Jose E. Labra *Universidad de Oviedo, Computer Science Department, Campus de los Catalanes, Oviedo, Spain. labra@uniovi.es*

ABSTRACT

Mailing list archives (i.e., the compilation of the messages posted up-to-now) are often published on the web and indexed by conventional search engines. They store a vast knowledge capital. However, the ability to automatically recognize and process the information is mostly lost at publishing time. As a result, the current mailing list archives are difficult to query and have a limited use. This paper describes an usage of the Semantic Web technologies in order to avoid the information loss and to allow new applications to exploit the information in a more powerful way.

KEYWORDS

Mailing list, Semantic web, SIOC, RDF, ontology

1. INTRODUCTION

Electronic mail (e-mail) remains one of the most popular applications of the Internet. Besides direct messaging between individuals, mailing lists exist as private or public forums for information exchange in communities with shared interests. Mailing list archives are compilations of the previously posted messages that are often converted into static HTML pages for their publication on the web. They represent a noteworthy portion of the contents that are indexed by web search engines, and they capture an impressive body of knowledge that, however, is difficult to locate and browse.

The root of these problems can be traced back to the translation procedure that is run to transform the e-mail messages into static HTML pages. This task is fulfilled by scripts that create an static HTML page for each message in the archive. In addition, some indexes (by date, by author, by thread) are generated and usually splitted by date ranges to avoid excessive growth.

On the one hand, this fixed structure reduces the flexibility when users browse the mailing list archives using their web browsers. On the other hand, some of the meta-data that were associated to each e-mail message are lost when the message is rendered as HTML for presentational purposes.

We propose to use an ontology and RDF (Resource Description Framework (Klyne 2004)) to publish the mailing list archives into the (Semantic) web, while retaining the meta-data that were present in the messages. Additionally, by doing so, the information could be merged and linked to other vocabularies, such as FOAF.

The rest of the paper is organized as follows: Section 2 introduces the SIOC ontology and our extensions to it, and then some software applications are described in Section 3. We close the paper with the conclusions and a discussion on future plans in Section 4.

2. SIOC

An ontology to capture the meta-data of a discussion forum, such as a mailing list, was clearly recognized as the first milestone to fulfill the purpose of the project. Fortunately, DERI Galway has developed SIOC (Semantically-Interlinked Online Communities, http://sioc-project.org/), an ontology that provides a vocabulary to interconnect different discussion methods such as blogs, web-based forums and mailing lists (Breslin 2005, Breslin 2006). Indeed, SIOC has a wider scope than just mailing lists, and groups all kinds of online discussion primitives in a generic sioc:Forum concept. Each forum represents an online community of people that share a common interest. The goal of SIOC is to interconnect these online communities. Other relevant concepts of the ontology are sioc:User and sioc:Post, which model respectively the members of the communities and the content they produce.

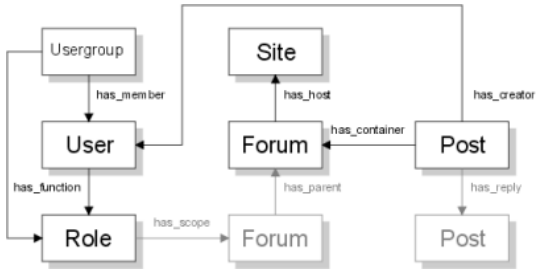

Figure 1. SIOC ontology terms

The SIOC ontology was designed to express the information contained both explicitly and implicitly in Internet discussion methods. Several software applications, usually deployed as plug-ins, are already available to export SIOC data from some popular blogging platforms and content management systems. The effort, however, is focused on web-based communities (weblogs, webforums), while little has been done so far to extend the coverage to legacy non-web communities, such as mailing lists and Usenet groups.

SIOC is specified in OWL, and their instances can be expressed in RDF. Therefore, they can be easily linked to other ontologies. The obvious choice here is FOAF (Brickley and Miller, 2005), which provides powerful means to describe the personal data of the members of a community.

2.1 Extending SIOC Ontology

SIOC is an almost perfect match for our purpose. Each mailing list becomes an instance of sioc:Forum, messages sent to the list become instances of sioc:Post (as well as their replies), and the people subscribed to the list are sioc:Users. The Dublin Core (Dublin Core Metadata Element Set, Version 1.1, 2006) vocabulary is used to capture meta-data such as the message date or title.

```
<rdf:RDF
  xmlns:dcterms='http://purl.org/dc/terms/'
  xmlns:sioc='http://rdfs.org/sioc/ns#'
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:swaml='http://swaml.berlios.de/ns/0.2#'
  xmlns:dc='http://purl.org/dc/elements/1.1/'
  xml:base='http://swaml.berlios.de/demo/'>
  <sioc:Post rdf:about="2006-Oct/post-50.rdf">
    <dc:title>SIOC properties cardinality</dc:title>
    <sioc:has_creator rdf:resource="subscribers.rdf#s4"/>
    <dcterms:created>Thu, 12 Oct 2006 23:59:26 +0200</dcterms:created>
    <sioc:content><!-- ommitted --></sioc:content>
    <sioc:has_reply rdf:resource="2006-Oct/post-51.rdf"/>
    <swaml:previousByDate rdf:resource="2006-Oct/post-49.rdf"/>
    <swaml:nextByDate rdf:resource="2006-Oct/post-51.rdf"/>
  </sioc:Post>
</rdf:RDF>
```

Figure 2. SIOC Post example in RDF/XML

However, additional object properties were required in order to retain the sequence of messages published in a mailing list. Thus, we extended the SIOC ontology with two properties defined in a separate namespace: swaml:previousByDate and swaml:nextByDate. Both properties are defined with sioc:Post as their domain and range. An RDF representation of a sample message is shown in Figure 2.

3. SOFTWARE TOOLS

The ontology itself provides no service to end users. Software tools are required, and we built two of them as part of this project[1]:

- SWAML is a non-interactive, command-line application whose main purpose is to translate mailboxes into sioc:Forum instances in RDF.
- Buxon is a graphical browser for sioc:Forum instances.

1 Our applications are available at http://swaml.berlios.de/

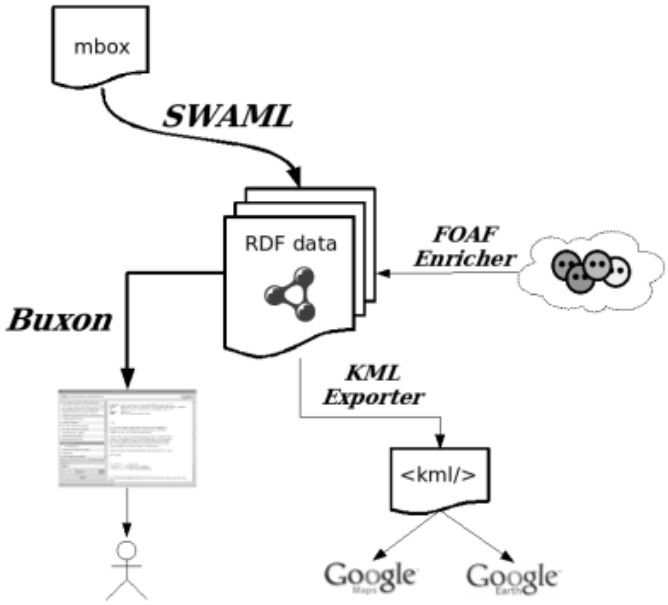

Figure 3. Buxon is an end-user application that consumes sioc:Forum instances, which in turn can be generated from mailboxes using SWAML.

Each tool has a precisely defined role, fulfilling the need to generate RDF data and to consume data, respectively, as depicted in Figure 3. The following paragraphs provide further detail on SWAML and Buxon.

3.1 SWAML

SWAML covers the data-generation phase, and it is intended to be used by mailing list administrators, who usually have access to the archives in raw format. The most popular format for mailing list archives is the "mailbox" (or "mbox"), as defined in RFC 4155 (Hall 2005). SWAML is essentially a mailbox parser implemented in Python. Its output is a number of SIOC instances (Forum, Posts and Users) in a set of RDF files. SWAML is a highly configurable, non-interactive application designed to be invoked by the system task scheduler.

Parsing the mailbox and rebuilding the discussion threads may be sometimes tricky. Although each mail message has a supposedly unique identifier in its header (Message-ID, defined by RFC 2822 (Resnick 2001)), in practice its uniqueness cannot be taken for granted. Actually, we have found some messages with repeated identifiers in some mailing lists, probably due to non-RFC compliant mail transport agents. Therefore, SWAML assumes that any reference to a message (such as those created by the In-Reply-To header) is in fact a reference to the most recent message with that ID in the mailbox (obviously, only previous messages are considered). Using this rule of thumb, SWAML builds an in-memory tree representation of the conversation threads, so sioc:Posts can be properly linked.

Actually, SWAML goes further than just a format-translation tool. A dedicated subroutine that runs as part of the batch execution, but may be also separately invoked on any sioc:Forum, tries to find a FOAF description for each sioc:User. To the best of our knowledge, there is not any web service to fetch FOAF descriptions from a given e-mail address, so we mocked it. Some of the authors of this paper are also currently working on a functional implementation of such a service as part of a different project.

The last step of the SWAML processing chain generates a KML (Ricket 2006) file that contains the geographical coordinates of the mailing list subscribers. The information is fetched from their FOAF descriptions, therefore it is only available for those subscribers whose FOAF description contains their coordinates using the basic geo vocabulary by Dan Brickley (Brickley 2006). Figure 4 depicts a graphical representation of the KML file for a sample mailing list.

Figure 4. Plotting the geographical coordinatesof the members of a mailing list using Google Maps.

3.2 Buxon

Buxon is a multi-platform desktop application written in PyGTK. It allows end users to browse the archives of mailing lists as if they were using their desktop mail application. Buxon takes the URI of a sioc:Forum instance (for example, a mailing list exported by SWAML, although any sioc:Forum instance is valid) and fetches the data, retrieving additional files if necessary. Then, it rebuilds the conversation structure and displays the familiar message thread list (see Figure 5).

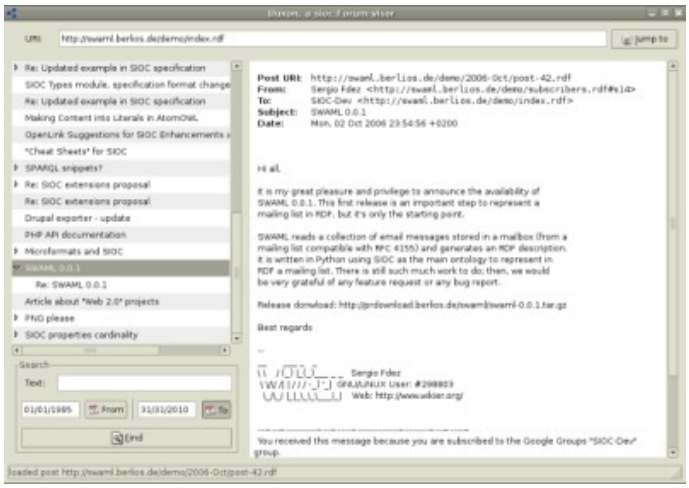

Figure 5. Buxon browsing SIOC-Dev mailing list.

Buxon also gives users the ability to query the messages, searching for terms or filtering the messages in a date range. All these queries are internally translated to SPARQL (Clark 2006) to be executed over the RDF graph, see Figure 6. Newer versions of Buxon can, at user's request, send the sioc:Forum URI to PingTheSemanticWeb.com, a social web service that tracks semantic web documents. That way, Buxon contributes to establish an infrastructure that lets people easily create, find and publish RDF documents.

```
PREFIX sioc: <http://rdfs.org/sioc/ns#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?title
FROM <http://swaml.berlios.de/demo/index.rdf>
WHERE
{
  ?x rdf:type sioc:Forum .
  ?x sioc:container_of ?message .
  ?message sioc:has_creator ?creator .
  ?creator sioc:name "Diego Berrueta" .
  ?message dc:title ?title
}
```

Figure 6. SPARQL query to extract all the posts sent by a given person to any sioc:Forum instance.

4. CONCLUSIONS AND FUTURE WORK

There is a lot of ongoing effort to translate data already reachable on the web into formats which are Semantic Web-friendly. Most of that work focuses on relational databases, micro-formats and web services. However, at the time of this writing and to the best of our knowledge, e-mail was almost excluded from the Semantic Web. This project, in combination

with the generic SIOC framework, fills this gap, conveniently providing an ontology and a parser to publish machine-readable versions of the archives of the countless mailing lists that exist on the Internet.

The SWAML project fulfills a much-needed requirement for the Semantic Web: to be able to refer to semantic versions of e-mail messages and their properties using resource URIs. By re-using the SIOC vocabulary for describing online discussions, SWAML allows any semantic web document (in particular, SIOC documents) to refer to e-mail messages from other discussions taking place on forums, blogs, etc., so that distributed conversations can occur across these discussion media. Also, by providing e-mail messages in SIOC format, SWAML is providing a rich source of data, namely mailing lists, for use in SIOC applications.

Some benefits arouse from the availability of these data. In the first place, data can be fetched by user applications to provide handy browsing through the archives of the mailing lists, providing features that exceed what is now offered by static HTML versions of the archives on the web.

Secondly, the crawlers of the web search engines can use the enhanced expressivity of the RDF data to refine search results. For instance, it becomes possible to filter out repeated messages, advance in the fight against spam, or introduce additional filter criteria in the search forms.

Another consequence of no lesser importance is that each e-mail message is assigned a URI that can be resolved to a machine-readable description of the message. This actually makes possible to link a message like any other web resource, and therefore enriches the expressivity of the web.

We are exploring some directions for future work. Some of them are:

- Integration of the SWAML process with popular HTML-based mailing list archivers, such as Hypermail or Pipermail, would be a giant push to speed up the adoption of SWAML. It is well known that one of the most awkward problems of any new technology is to gain a critical mass of users. The semantic web is not an exception. A good recipe to tackle this problem is to integrate the new technology into old tools, making a smooth transition without requiring any extra effort from users. Merging the SWAML process into the batch flow of tools such as Hypermail would allow to generate both HTML and RDF versions of the archives. Those could reside side-by-side on the web server, even sharing the same URI by means of content-negotiation (Miles 2006).
- Actually, integration could be pushed further away through RDFa (Birbeck 2006), embedding the RDF content into the XHTML documents.
- So far, no semantic annotation relative to the meaning of the messages is considered. Obviously, such information can not be automatically derived from a RFC 4155-compliant mailbox. However, it is conceivable that it can be added by other means, such as social tagging using folksonomies, or parsing the RDFa that may exist in the e-mail messages that are sent in XHTML format. The inherent community-based nature of mailing lists can be exploited to build recommendation systems (Celma 2006).
- The meta-data extracted from a mailing list archive can grow quite huge. Even if the body of the messages is omitted, the RDF/XML meta-data of a mailing list containing 1,000 messages may have a size of 4 MBytes, with a linear growth. It is not uncommon for a busy mailing list to generate such volume of messages monthly. Hence, it becomes imperative to provide a mechanism to fragmentate the dataset. The SWAML process splits each message in a separate RDF document, but this arbitrary decision clearly does not fit every application. A much better solution would be to create an easy-to-deploy SPARQL endpoint (Clark 2006), effectively translating the decision on how to partition the data to

the final application (Pan 2006).

- It is not always possible to obtain a mailbox file for a mailing list. For these cases, an alternative is envisaged: a high-capacity mail account can be subscribed to the mailing list with the unique purpose of collecting and storing the messages. A simple extension to SWAML that makes it possible to read the contents of a GMail account has been developed.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Dr. John Breslin and Uldis Bojars from DERI Galway, whose support and contributions have been of great help to this project. Also to Ignacio Barrientos by his contribution packaging the project for Debian GNU/Linux.

REFERENCES

Birbeck, M. et al, 2006. RDFa Syntax, a collection of attributes for layering RDF on XML languages, Technical report, W3C.

Breslin, J. et al, 2006. SIOC: an approach to connect web-based communities. *International Journal of Web Based Communities*, Vol. 2, No. 2, pp 133-142.

Breslin, J. et al, 2005. Towards Semantically-Interlinked Online Communities. *Proceedings of the 2nd European Semantic Web Conference, ESWC 2005*, Heraklion, Crete, Greece.

Brickley, D., 2006. Basic geo (WGS84 lat/long) vocabulary, Technical report, W3C Informal Note.

Brickley, D. & Miller, L., 2005. FOAF Vocabulary Specification, Technical report.

Celma, O., 2006. Foafing the music: Bridging the semantic gap in music recommendation. *Proceedings of the 5th International Semantic Web Conference*, Athens, USA.

Clark, K. G., 2006. SPARQL protocol for RDF, Technical report, W3C Candidate Recommendation.

Dublin Core Metadata Element Set, Version 1.1, 2006. Technical report.

Hall, E., 2005. RFC 4155 - the application/mbox media type, Technical report, The Internet Society.

Klyne, G. and Carroll, J. J., 2004. Resource Description Framework (RDF): Concepts and abstract syntax, Technical report, W3C Recommendation.

Miles, A. et al, 2006. Best practice recipes for publishing RDF vocabularies, Technical report, W3C Working Draft.

Pan, Z. et al 2006. An investigation into the feasibility of the semantic web, Technical Report LU-CSE-06-025, Dept. of Computer Science and Engineering, Lehigh University.

Resnick, P., 2001. RFC 2822 - internet message format, Technical report, The Internet Society.

Ricket, D, 2006. Google Maps and Google Earth integration using KML, in American Geophysical Union 2006 Fall Meeting.

A CONVERSION PROCESS FROM FLICKR TAGS TO RDF DESCRIPTIONS

Mohamed Zied MAALA
France Telecom R&D
38-40 rue du General Leclerc, 92130 Issy-les-Moulineaux Cedex 9, France
zied.maala@orange-ftgroup.com

Alexandre DELTEIL
France Telecom R&D
38-40 rue du General Leclerc, 92130 Issy-les-Moulineaux Cedex 9, France
alexandre.delteil@orange-ftgroup.com

Ahmed AZOUGH
France Telecom R&D
38-40 rue du General Leclerc, 92130 Issy-les-Moulineaux Cedex 9, France
ahmed.azough@orange-ftgroup.com

ABSTRACT

The recent evolution of the Web, now designated by the term Web 2.0, has seen the appearance of a huge number of resources created and annotated by users. However the annotations consist only in simple tags that are gathered in unstructured sets called folksonomies. The use of more complex languages to annotate resources and to define semantics according to the vision of the Semantic Web, would improve the understanding by machines and programs, like search engines, of what is on the Web. Indeed tags expressivity is very low compared to the representation standards of the Semantic Web, like RDF and OWL. But users appear to be still reluctant to annotate resources with RDF, and it should be recognized that Semantic Web, contrary to Web 2.0, is still not a reality of today's Web. One way to take advantage of Semantic Web capabilities right now, without waiting for a change of the annotation usages, would be to be able to generate RDF annotations from tags. As a first step toward this direction, this paper presents a tentative to automatically convert a set of tags into a RDF description in the context of photos on Flickr. Such a method exploits some specificity of tags used on Flickr, some basic natural language processing tools and some semantic resources, in order to relate semantically tags describing a given photo and build a pertinent RDF annotation for this photo.

KEYWORDS

Web 2.0, tags, RDF, annotation generation.

1. INTRODUCTION

Web 2.0 and Semantic Web are two trends influencing the evolution of the Web since several years. Web 2.0 consists in a greater collective content creation and a larger social interaction between users. A huge number of resources have been created by users and annotated by them. This is a major change compared to the original web, where collective creation was much less developed. Many resource repositories like Wikipedia [1], Del.icio.us [2] and Flickr [3] have appeared and gather millions of user created pages, bookmarks and photos. However up to now the annotations made by the users on these resources consist only in simple tags, that are gathered in unstructured sets called folksonomies and thus do not convey a formally defined

semantics. Therefore they do help to improve queries on the Web, but not so much, since resources will be found only if the query syntactically matches a tag they are annotated by.

Contrary to Web 2.0, Semantic Web is still a vision and not yet a reality. It is based on the idea that describing resources with symbolic annotations (using vocabularies defined in formal ontologies) will enable machines and tools to understand their semantics and will improve the pertinence of tasks such as query answering.

This paper focuses in the study of the plateform Flickr [3], a photo sharing website and web services suite. Flickr [3] was developed by Ludicorp, a Vancouver, Canada-based company founded in 2002. Ludicorp launched Flickr in February 2004. In March 2005, Yahoo! Inc. acquired Ludicorp and Flickr. Flickr allows photo submitters to categorize their images by use of keywords "tags" (a form of metadata), which allow searchers to easily find images concerning a certain topic such as place name or subject matter.

Flickr [3] provides rapid access to images tagged with the most popular keywords. Flickr also allows users to categorize their photos into "sets", or groups of photos that fall under the same heading. However, sets are more flexible than the traditional folder-based method of organizing files, as one photo can belong to many sets, or one set, or none at all (the concept is directly analogous to the better known "labels" in Google's Gmail). Flickr's "sets", then, represent a form of categorical metadata rather than a physical hierarchy.

This paper interests more exactly to the study of Flickr tags and present a new method to convert Flickr [3] tags describing a picture into RDF annotations describing it semantically. This method can be viewed as the first step enabling to transform resources described using tags to a semantic description describing the same resources or can be viewed too as a first bridge between the web 2.0 and the semantic web. This method is based on linguistic rules, on natural language treatment, on integrating some human knowledge to be able to provide semantic description for pictures from tags. To the best of our knowledge, few works exist enabling conversion from tags to semantic description. One work [4] exists that "converts" Flickr tags to RDF descriptions, but gives bad results because Flickr tags are transformed into RDF topics in a fully syntactic way without extracting the semantic of the tags. Our method helps a user to understand picture tags and to found relationships between them.

This paper contains five sections. Section 2 presents the different ways of tagging pictures used in Flickr. Section 3 introduces our conversion method from tags to RDF semantic description. Section 4 describes related work and compares our method with existing approaches.

2. SURVEY ON WAYS OF TAGGING PICTURES IN FLICKR

Before conceiving a method to generate a RDF description from tags on Flickr, it is useful to know the specificities of photo tag annotations. Therefore we have attempted to analyse the different ways users exploit Flickr annotation capabilities in order to tag photos.

2.1 Tagging habits

The following tagging habits can be distinguished:

- Very few tags: unfortunately too many photos contain no tag at all or very few tags (one or two such in figure 1). In this case, it is impossible or very difficult to generate a RDF description.
- Sentence tagging: users can use quotes to enter a full sentence as a tag such as in figure 2 (in case no quotes are used, space is understood by Flickr as a separator between tags).
- Vertical sentence tagging: it is the same case as the previous one, but users forgot to (or intentionally did not) put the sentence between quotes. Thus the sentence can be read vertically, because Flickr has understood each space separated word to be a different tag (such as in figure 3).
- Too many tags: contrary to the previous case, the information attached to the photo is very rich (as in figure 2) and describes many different aspects (content, location ...). The difficulty for

generating a RDF description lies in finding the relevant associations between the tags (for instance which noun is subject of which verb).

- Non-sense tags: these tags correspond to something not understandable for a human being not knowing the annotator universe of thinking such as in figure 2 (for instance the tag *noneof100#2*). It could for instance be a nickname of some people on the photo, or of a location...
- Space free tagging: the users write a sentence by concatenating words in order to put the whole sentence on the same line ; for example in figure 2 a user has written the tag "I love nature". These users may not be aware of the possibility of using quotes.
- Collective tagging: due to the interface Flickr provides (see figure 4), it is possible to tag several photos concurrently. Therefore it sometimes happens that a photo is described with a tag that does not apply directly to it but to a photo that has been uploaded at the same time. The photo the tag applies to belongs to the previous five or next five photos of the current photo in the "photostream" (as six photos can be concurrently tagged).

Fig 1. Use of few tags
Tags{Hawai, Tourist}

Fig 2. Use of sentence as tags
Tags{Paya Lake, Makra top, Kaghan valley, nature, water, I love nature, wow, noneof100#2, top-v111, top-v1111, deleteme, saveme, saveme2, saveme3, deleteme2, deleteme3, saveme4, saveme5, deleme4, saveme6, deleteme4, deleteme5, deleteme6, deleteme7, deleteme8, saveme7, deleteme9, saveme8, deleteme10, saveme9, Most, bravo, Big Fave, Outstanding shots }

Fig 3. Tag with vertical sentence
Tags{ Here, some, more, photos, off, Hudson, River, New, Jersey }

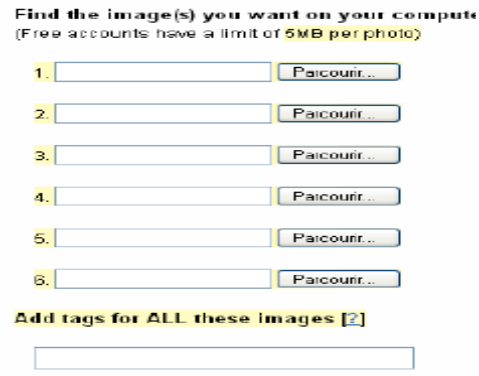

Fig 4. Collective tagging due to pictures upload user interface

Moreover many tags contain typing errors, due perhaps to a too high typing speed or to a lack of knowledge of correct typing. The figure 5 presents a histogram of photo tags number on a sample size of one thousand photos. This figure shows more precisely the distribution of the number of keywords.

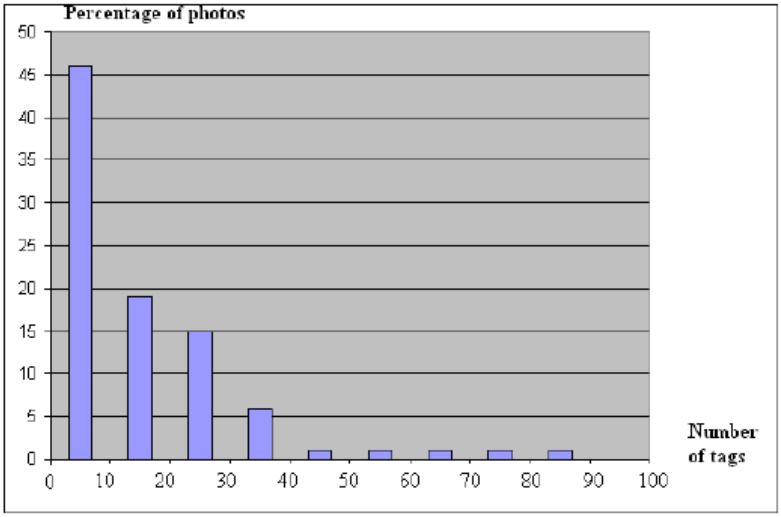

Fig 5. Distribution of the number of keywords

2.2 Flickr interface

To know which tags are the most employed by users and their links with pictures, we studied popular tags presented in figure 6. On 145 popular tags, 10% describe a celebration (birthday, Christmas), 13% are related to a date (June, July...), and 62 tags (about 42% of the total) express places, on 20 tags are names of common places (mountain, house, garden,...) and 42 tags (about 28% of the total) are name of countries or towns (Canada, Japan, Paris,...). Among the popular tags, some are related to the camera (Nikon, cameraphone ...).

Fig 6. Popular tags

Table 1 shows percentage of photos by category.

Table 1. Photo percentage by category

Tag category		Percentage of photos	
Place	Country	38	71
	Landscape	42	
	Building	10	
Time	Year/Season	20	20
	Month/Day	5	
Event		11	
Name		35	
Camera		17	
Action		53	
Non sense		49	

The next section presents more precisely what the tags describe exactly about the photo.

2.3 Tagging content

An analysis of a sample of one thousand photos shows that the tags can be clustered in the following groups:

- Place: the location can be described at very different levels of granularity. At the largest level of granularity, the continent, the country, the region, the city, a mountain range . . . are found frequently. At a smaller level of granularity, description of the building or the immediate natural site the photo has been taken in can be found: a building, a university, a house, a beach… Finally at the smallest level of granularity, there can be a description of a room or a piece of furniture: bed, chair . . .
- Time: the time can also be described at different levels of granularity. The year, the season and the month are the most frequently found. The exact day is much less frequent. Some times of the day are (sunrise, sunset…).
- Event: the holy days (Christmas . . .), the birthdays, the weddings . . .
- Name: people names (Emma, Jean . . .), nicknames. . .
- Camera: many tags indicate the make or the model of the camera (Nokia, Canon . . .), the colors (black & white ...), artistic judgments on the photo . . .

This knowledge of the way people tag photos on Flickr gives an indication on the natural language processing tools and semantic resources that are needed in order to be able to transform a set of tags into an RDF annotation. The process of automatically generating RDF annotations is now described in next section.

3. CONVERSION PROCESS OF FLICKR'S TAGS TO RDF ANNOTATION

This section introduces a method to convert tags describing a photo into a RDF annotation. This method can be viewed as a first tentative to transform web 2.0 annotations into semantic web annotations. The problem can also be viewed as transforming a bag of tags into a relational description. This method mainly relies on detecting the category each tag belongs to, among a set of six categories (location, time, event, people, camera, activity). Using this set of categorized tags, it then tries to identify the possible arguments of verbs (verbs are in the category denoting activity) in infinitive or present participle form. This method thus

applies only on photos described by tags when some of them are verbs. In the next sections, the components needed in the conversion process are described one by one.

3.1 CONVERSION PROCESS COMPONENTS

Automatic conversion from photo tags to RDF annotations is a difficult task. This process essentially requires several components: some basic natural language processing tools (mainly a stemmer), and semantic resources like Wordnet, semantic nets and specialized databases containing knowledge on specific subjects (for instance locations, cameras ...).

3.1.1 Wordnet

WordNet [7] is a semantic lexicon for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. The database can also be browsed online. WordNet [7] was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. As of 2006, the database contains about 150000 words organized in over 115000 synsets for a total of 207000 wordsense pairs, 11488 verbs, 22141 adjectives, 4601 adverbs. WordNet [7] distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. Every synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning, such as “car pool”); different senses of a word are in different synsets. The meaning of the synsets is further clarified with short defining glosses (Definitions and/or example sentences). For example, the noun vacation has two senses. The first sense of the word vacation is given by a synonym holiday and the definition: leisure time away from work devoted to rest or pleasure. The second sense of the word vacation is given by the definition: the act of making something legally void. Most synsets are connected to other synsets via a number of semantic relations.

These relations are based on the type of word, and include:

- Nouns
  – Hypernyms: Y is a hypernym of X if every X is a (kind of) Y – hyponyms: Y is a hyponym of X if every Y is a (kind of) X
  – Coordinate terms: Y is a coordinate term of X if X and Y share a hypernym
  – Holonym: Y is a holonym of X if X is a part of Y
  – Meronym: Y is a meronym of X if Y is a part of X
- Verbs
  – Hypernym: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (travel to movement)
  – Troponym: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (lisp to talk)
  – Entailment: the verb Y is entailed by X if by doing X you must be doing Y (sleeping by snoring)
  – Coordinate terms: those verbs sharing a common hypernym
- Adjectives
  – Related nouns
  – Participle of verb
- Adverbs
  – root adjectives While semantic relations apply to all members of a synset because they share a meaning but are all mutually synonyms, words can also be connected to other words through lexical relations, including synonyms, antonyms (opposites of each other) and derivationally related, as well. WordNet [7] also provides the polysemy count of a word: the number of synsets that contain the word. If a word participates in several synsets (i.e. has several senses), then typically some senses are much more common than others.

3.1.2 Knowledge resources

As it has already been explained in 2.3, most of Flickr photos are described by tags that denote:

- Places: continents, countries, cities, natural environment, objects on which (or in which) people can stand (buildings, furniture ...)
- Time: years, seasons, days...
- Events: Christmas, birthday...
- Names: Emma, Jean, nicknames...
- Cameras : Nokia, Canon, colors...

In order to be able to understand the meaning of these tags and correctly build a RDF annotation, some semantic resources are needed. For each tag category described above, the resources have been either created or crawled from the web and sometimes completed.

- Places: two place resources are used, a database containing geographical locations (for instance Los Angeles is in California which is in the US which is in America) and an ontology of things where people can be (for instance people can be at a table, which can be inside a house, which can be inside a city; or people can be in a car, that can be on a road, that can be in a state, ...).

  – For the first one, we crawled several websites (like for instance Yahoo! Meteo) to obtain lists of cities, with the countries and continents in which they are located.
  – For the second one, we had to complete Wordnet in order to be able to infer which kind of things could be a location for people. This consisted in adding about 200 location relations (meaning “can be located in”).
- Time: there are not so many concepts for denoting time; we completed Wordnet and obtained an ontology of about 50 concepts (containing seasons, days, months, moments of the day ...)
- Events: as for time, we completed Wordnet and obtained an ontology of about 50 concepts denoting events (birthday, wedding, vacation, holy days ...)
- Cameras: we gathered a set of makes and models by crawling online shopping websites (for new and used products).

The method presented in this paper tries to convert a set of tags into a RDF annotation. The RDF language is thus presented in the following section.

3.1.3 RDF

Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata model but which has come to be used as a general method of modeling knowledge, through a variety of syntax formats. The RDF metadata model is based upon the idea of making statements about resources in the form of subject-predicate-object expressions, called triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. For example, one way to represent the notion ”The sky has the color blue” in RDF is as a triple of specially formatted strings: a subject denoting ”the sky”, a predicate denoting ”has the color”, and an object denoting ”blue”.

Below, a RDF ressource description introducing a “Person” whose name is “Emma” is presented:

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://www.emma.htm">
<dc:title>Emma</dc:title>
<dc:publisher>PersonalPage</dc:publisher>
<foaf:primaryTopic>
<foaf:Person>
<foaf:name>Emma</foaf:name>
</foaf:Person>
</foaf:primaryTopic>
```

</rdf:Description>
</rdf:RDF>

3.2 CONVERSION PROCESS DESCRIPTION

This section describes the method enabling to generate a RDF description from a set of Flickr tags. Figure 7 shows the main components used in the process. It takes in input all the tags describing a photo and returns in output a RDF description of the photo. The semantic relations in, at, by, event, shot − by, describes, agent and object are introduced to form the resulting RDF annotation. The different steps are then the following (the photo is denoted by r):

- A stemmer enables to transform a tag into its non inflectional form,
- Using the semantic resources, each tag is then categorized in one of the six categories (location, time, event, people, camera, activity),
- All tags grouped in the location category are ordered from the smallest to the largest, say $l_1 \leq l_2 \leq \ldots \leq l_n$). The generated triples are: *(r, in, l_1), (l_1, in, l_2). . . (l_{n-1}, in, l_n)*.
- Similarly all tags grouped in the tag category are ordered from the smallest to the largest, say $t_1 \leq t_2 \leq \ldots \leq t_n$). The generated triples are: *(r, at, t_1), (t_1, at, t_2). . . (t_{n-1}, at, t_n)*.
- For each event e a triple (r, event, e) is created,
- For each camera c a triple (r, shot − by, c) is created,
- For each verb *v* in the activity category, the corresponding signature is retrieved from Wordnet, say x → y.

For each tag ***a*** of type ***x*** and each tag ***y*** of type ***y***, the triple *(r, describes, v)*, *(v, agent, x)* and *(v, object, y)* are added.

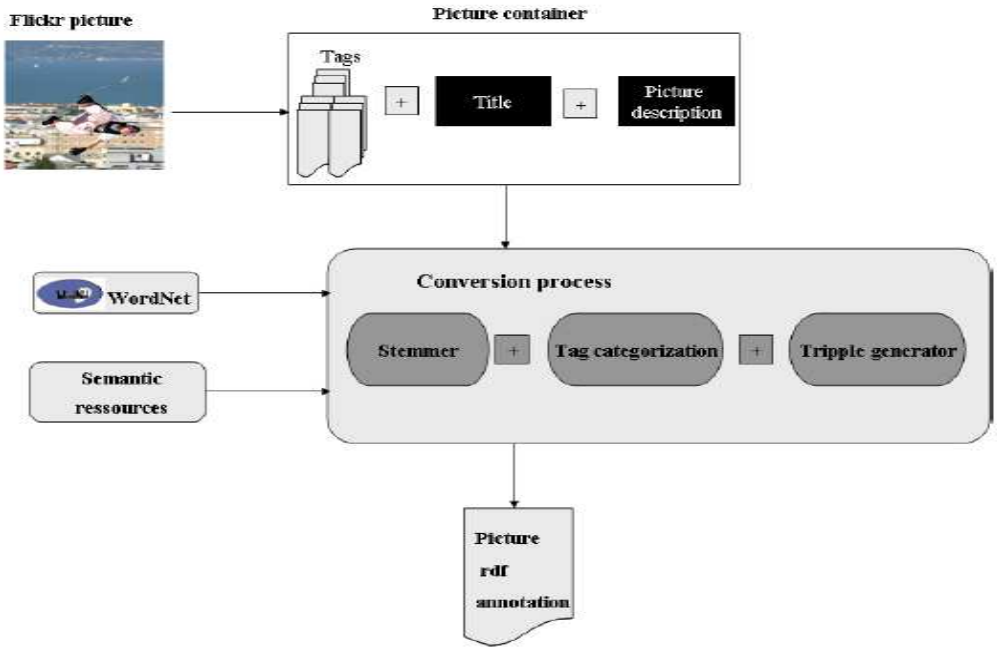

Fig 7. Conversion process components

For instance, tags describing the photo represented in figure 8 will lead to the annotation: *(r, in, NewHartford), (NewHartford, in, Connecticut), (r, describes, ski), (ski, agent, JeffG)*.

Fig 8. Example 1
Tags{Connecticut, skiing, New Hartford, Jeff G}

The tags describing the photo represented in figure 7 will lead to the annotation: *(r, in, Austria), (r, event, honeymoon), (r, at, January), (January, at, 2007)*.

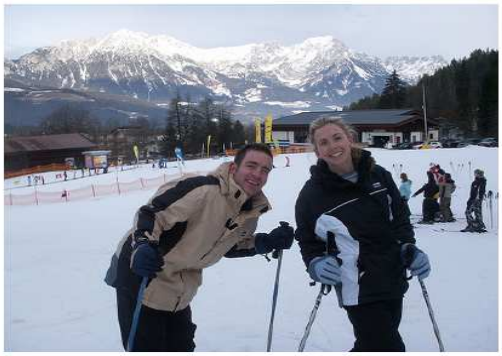

Fig 9. Example 2
Tags{Soll, Austria, Skiing, Honeymoon, January, 2007}

The tags describing the photo represented in figure 10 will lead to the annotation: *(r, in, queensland), (queensland, in, Australia), (r, describes, fly), (fly, agent, birdie)*.

Fig 10. Example 3
Tags{pelican, birdie, queensland, flying, Australia, iansand, waterfowl, peopleplacesevents}

The tags describing the photo represented in figure 11 will lead to the annotation: *(r, in, Liege), (Liege, in, Belgium), (r, describes, drive), (drive, agent, Gaelle)*.

Fig 11. Example 4
Tags{motion, Belgium, liege, drive, gaelle, starlet}

The tags not recognized by the different semantic resources (for instance *iansand, peopleplacesevents…*) are ignored during the conversion process.

4. RELATED WORK

A close research problem to ours is that concerning semiautomatic generation of annotations. [6] explain how, based on KA community initiative (Knowledge Annotation initiative of the Knowledge Acquisition community), an ergonomic and knowledge base-supported annotation tool was developed, and how this tool was extended with mechanisms that semi-automatically propose new annotations to the user. Supporting the evolving nature of semantic content, authors describe their idea of evolving Ontologies supporting semantic annotation; they conclude that semantic annotation and ontology engineering must be considered as a cyclic process. Although this work is important, some issues remain unsolved in this paper. Authors mentioned that an integrated system of annotation and ontology construction combining knowledge base-supported, ergonomic annotation, with an environment and methods for ontology engineering and learning from text supporting evolving Ontologies should be build. Furthermore, ergonomic, ontology and semiautomatic suggestion of the system should be evaluated. In addition annotated facts are not reusable since the system didn't support the RDF standard for representing metadata on the web.

Another work in the same domain it the one done by [8]. In this work a framework, S-CREAM, was developed to that allows for creation of metadata and is trainable for a specific domain. It supports the semi-automatic annotation of web pages based on the information extraction component Amilcare. It extracts knowledge structure from web pages through the use of knowledge extraction rules. These rules are the result of a learning-cycle based on already annotated pages. Authors are further investigating how different tools may be brought together, e.g. to allow for the creation of relational metadata in PDF, SVG, or SMIL.

Not very far from this [5] treat the generation of Ontologies. [5] present a comprehensive architecture and generic method for semi-automatic ontology acquisition from given intranet resources. A new approach for supporting the overall process of engineering Ontologies from text is described. Based on a given core ontology extended with domain specific concepts, the resulting ontology is restricted to a specific application using a corpus-based mechanism for ontology pruning. On top of the ontology two approaches supporting the difficult task of determining non-taxonomic conceptual relationships are applied. To complete this work several techniques for evaluating the acquired ontology should be developed. Also it should be elaborated how the results of different learning algorithms will have to be assessed and combined in the multi-strategy learning set newly introduced by the authors.

Some works were done on the Conversion of WordNet to a standard RDF/OWL representation. [9] presents an overview of the work in progress at the W3C to produce a standard conversion of WordNet to the RDF/OWL representation language in use in the Semantic Web community. The paper explains the steps involved in the conversion and details design decisions such as the composition of the class hierarchy and properties, the addition of suitable OWL semantics and the chosen format of the URIs. Some issues remain open like supporting different versions of WordNet in RDF/OWL and defining the relationship between

them. Furthermore, the integration of WordNet with sources in other languages is not solved. Most of existing works provide a semi-automatic generation of annotations.

[4] is a tool that converts automatically Flickr tags to RDF. However it does not provide a really semantic description of photos but it rather syntactically translates each tag in a separate RDF triple.

5. CONCLUSION AND FUTURE WORK

This paper has presented a conversion process from Flickr's photo tags to RDF annotations, thus leading to a first bridge between Web 2.0 and Semantic Web. Before conceiving this method, the ways people used to tag photos on Flickr were analyzed. It has shown that people mainly employed six categories of tags, each one denoting a certain aspect of the photo: location, time, event, people, camera, and activity. For each one of these categories, semantic resources have been either reused and completed (like Wordnet) or crawled from the web (like camera and location databases). Using these semantic resources, the method presented in this paper tries to identify the category of each tag. It then uses the signatures of verbs (tags of category activity) in Wordnet to associate a verb with its subject and complement and thus to build a RDF triple. Other triples are built by using tags of other categories, for instance by linking the photo with the smallest location as well as a location with a more general location. This method gives its best results for photos containing in their tags verbs, as these tags will provide the RDF relations that are the less common and thus the most interesting. Future work will try to take advantage of the presence of other information (the title and the legend of the photo) to improve the understanding of what the photo is about and to generate a RDF description that is more accurate.

6. REFERENCES

[1] http://www.wikipedia.org/.

[2] http://del.icio.us/.

[3] http://www.flickr.com/.

[4] www.kanzaki.com/works/2005/imgdsc/flickr2rdf.

[5] A method for semi-automatic ontology acquisition from a corporate intranet. In Proceedings of EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000, number 1937 in Springer Lecture Notes in Artificial Intelligence (LNAI), 2000.

[6] M. Erdmann, A. Maedche, H. Schnurr, and S. Staab, 2000, From manual to semi-automatic semantic annotation: About ontology-based text annotation tools.

[7] C. Fellbaum, 1998, Wordnet: an electronic lexical database, *Language, Speech and Communication, Cambridge*.

[8] S. Handschuh, S. Staab, and F. Ciravegna, 2002, S-cream — semi-automatic creation of metadata. *In 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*.

[9] M. van Assem, A. Gangemi, and G. Schreiber, 2006, Conversion of WordNet to a standard RDF/OWL representation. *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*.

Workshop on

Ontology Evolution (OnE 2007)

26th of April 2007,
Poznań, Poland

Workshop Program Committee

Sven Casteleyn, Vrije Universiteit Brussel, Belgium
Emanuele Della Valle, CEFRIEL, Italy
Giorgos Flouris, ISTI, CNR, Italy
Peter Haase, AIFB - University of Karlsruhe, Germany
Zhisheng Huang, Vrije Universiteit Amsterdam, The Netherlands
Michel Klein, Vrije Universiteit Amsterdam, The Netherlands
Jacek Kopecky, University of Innsbruck, Austria
Oliver Kutz, The University of Manchester, UK
Peter Plessers, Vrije Universiteit Brussel, Belgium
Krzysztof Wecel, Poznan University of Economics, Poland

On the Evolution of Ontological Signatures

Giorgos Flouris

ISTI-CNR
Via G. Moruzzi, 1,
56124, Pisa, Italy

flouris@isti.cnr.it

ABSTRACT

During ontology evolution, we are often faced with operations requiring the addition/removal of some ontological element (e.g., a concept) to/from the signature. Such operations deal with the ontological signature and are fundamentally different from operations that deal with the axiomatic part of the ontology, because they don't affect our knowledge on the domain but the non-logical symbols of the logic used to represent our knowledge on the domain. The consequences of this observation have been generally disregarded in the relevant literature. This paper attempts to fill this gap by introducing the concept of "change levels" and discussing the issues emerging from the different nature of the two types of operations. Furthermore, two alternative formalizations are described, which allow both types of operations to be represented at the same level, and, consequently, be considered of the same type.

1. INTRODUCTION

An *ontology* can be defined as a pair **<*S*,*A*>**, where ***S*** is the *vocabulary* (or *signature*) of the ontology and ***A*** is *the set of ontological axioms* [11]. The signature is usually modeled as a simple set containing the names of all concepts, properties or individuals that are relevant to the domain of discourse, while the axioms specify the intended interpretation of these symbols (names) in the given domain of discourse.

Given this definition, it seems reasonable that changes upon ontologies should affect both the signature and the ontological axioms. Indeed, ontology evolution has traditionally dealt with both types of changes and many works on ontology evolution handle both types of changes in a similar manner (e.g., [9], [16], [19], [20]).

However, the admittance of such operations is unique in the ontology evolution context; in the main research area studying changes upon a corpus of knowledge, namely *belief revision* [8] (also known as *belief change*), the signature (called *language* in that context) is considered static, so these types of changes are not considered.

As a result, the incorporation of signature changes in ontology evolution disallows the use of many of the formal tools provided by the related field of belief revision [7]. Thus, it is not surprising that many of the recent works in ontology evolution, especially the more theoretically-minded ones, do not consider such changes (e.g., [10], [13], [17], [18]).

In this paper we argue that treating both types of changes in the same manner is rather problematic from a methodological point of view, because the axioms and the signature each constitute a fundamentally different "knowledge level", so their respective change operations should be handled separately. This intuition is captured by introducing the concept of "change levels" (section 2), which allows the formal study of the two types of operations. In addition, two representation methodologies are introduced, which allow the incorporation of the signature information into the axiomatic part of the ontology, thus allowing a homogeneous treatment of both operation types (section 3).

Even though most of the results presented in this paper are applicable to many different kinds of representation formalisms and contexts, our focus will be the ontological context; standard logical Knowledge Bases (KBs) will be used for comparison. It will be assumed, for simplicity, that ontologies are represented using some Description Logic (DL) and logical KBs are represented using First-Order Logic (FOL), so the reader is assumed to have some basic familiarity with DLs [1] and FOL [4].

2. CHANGE LEVELS

2.1 Components of the Symbol Level

In his seminal work [15], Newell identified two major levels in every system (knowledge representation or other). The first, the *knowledge level*, contains all the abstractions that are used to describe a system's behavior and is independent of any implementation peculiarities; the second, the *symbol level*, contains the mechanisms (formalisms) that allow the system to operate.

Here, we focus on the symbol level; in the context of Knowledge Representation (KR), this level contains the axioms or formulas that describe system's knowledge (i.e., the KB). A KB is based on some logical formalism and uses various non-logical symbols (names) representing concepts, properties, predicates etc, depending on the context. The role of the KB is to capture the intended interpretation of the non-logical symbols in the domain of discourse using logical formulas; the semantics, syntax etc of these formulas is provided by the underlying logical formalism. This analysis motivates viewing the symbol level as being structured from these three clearly defined, but interrelated, components (levels): the *logic*, the *language* and the *knowledge base* (see table 1).

The first level (logic) is used to describe the logical elements (symbols) of the formalism that is used to represent our knowledge (e.g., connectives). Moreover, the semantics, syntax and inference mechanisms of the logic are all included in the logic level. In the ontological context, this level consists of the formal definition of the formalism used to formulate the axioms (e.g., DL [1], OWL [3], RDF [14] etc).

In the second level (language), the non-logical elements that are relevant to the domain are identified. These non-logical elements are, essentially, the (intuitive) names that we give to the various

-Table 1. Levels of Knowledge Representation (Components of the Symbol Level)

Components of the Symbol Level	Example: Knowledge Bases and Standard Logics	Example: Ontologies and Description Logics
Level 1: *Logic* Logical symbols, semantics, syntax, inference mechanism	FOL First-order connectives (e.g., ∀, ∃, ∧, …) Semantics of FOL Syntactical rules for FOL FOL inference rules	*ALC* *ALC* operators and connectives (e.g., ⊓, ¬, ⊑, …) *ALC* semantics Syntactical rules for *ALC* *ALC* inference rules
Level 2: *Language* Vocabulary and terminology of the domain	Non-logical symbols (names of predicates, functions etc)	Signature structure (names of concepts, properties etc)
Level 3: *Knowledge Base* Axioms, propositions	KB (set of FOL formulas)	Ontological axioms (set of *ALC* axioms)

relevant concepts, properties, predicates etc. This level corresponds to the signature of an ontology.

The third level (KB) is the actual embodiment of our knowledge on the domain. This level describes the interrelationships between the various elements of the language level; the types (and the semantics) of the allowed interrelationships are determined by the logic level. Obviously, the KB-level cannot be defined without an explicit and detailed description of the other two levels. In the ontological context, it is represented by the ontological axioms.

2.2 Language-level and KB-level Changes

The discrimination of the various components of the symbol level motivates a similar discrimination between the various types of changes on the basis of the component of the symbol level that they affect (see table 2).

In particular, the term *KB-level change* will be used to refer to change operations that directly affect the KB level of a KR system. Examples of KB-level changes in ontology evolution are the addition or removal of an IsA or a restriction upon the range of a property. An example of a KB-level operation in the standard logical setting (belief change) is contraction.

The term *language-level change* will be used to refer to change operators that directly affect the second level in table 1. Examples of language-level changes are the addition or removal of concepts, roles or individuals from the signature. In the standard logical setting, such operators are not considered, because the language is assumed to be static.

In principle, it is also possible to define *logic-level* changes, referring to changes that directly affect the logic itself. An example of such a change would be "remove the operator ⊓ from the underlying DL". However, the underlying logical formalism is usually considered static: neither belief revision nor ontology evolution deal with such operations.

Notice that the word "directly" is necessary in these definitions, because it is possible for a change to have side-effects affecting different levels. This is true because the three levels are not stand-alone entities but affect and depend on each other.

In particular, the removal of an element from the signature may have side-effects on the axiomatic part of the ontology; for example, if we are asked to remove a concept, then all axioms that refer to this concept (e.g., classification axioms) must be removed or otherwise amended so as not to involve the removed concept; all such amendments are KB-level changes.

A similar situation may occur when adding axioms; for example, if we are asked to add an IsA relation between concepts A and B and B does not exist in the ontological signature, then it should either be added (as a concept), or the operation should be rejected. In this case, a KB-level change may have a language-level side-effect.

On the other hand, removing an axiom from an ontology cannot cause any language-level changes. Some would argue that if, after the removal of an axiom, nothing is known regarding a certain element (e.g., a concept), then this element should be removed. This viewpoint is rather problematic. The fact that no interesting information regarding an element can be inferred from an ontology means that nothing is really known about this particular

Table 2. Change Levels and Their Support in Belief Change and Ontology Evolution

Change Levels	Belief Change	Ontology Evolution
Level 1: *Logic* Logic-level changes (affect the logic)	Does not support changes at this level	Does not support changes at this level
Level 2: *Language* Language-level changes (affect the language)	Does not support changes at this level	Supports changes at this level; changes may have side-effects in level 3
Level 3: *Knowledge Base* KB-level changes (affect the KB)	Supports changes at this level; changes cannot affect other levels; if they do, they are rejected as non-valid	Supports changes at this level; changes may have side-effects in level 2

element (yet). On the other hand, removing an element from the ontological signature implies that this element is irrelevant to the conceptualization of the domain described by the ontology; this statement is fundamentally different from the previous one. Therefore, it can be argued that, if the ontology engineer wishes to state that a particular element is irrelevant to the ontology, he should do so explicitly, by removing the element from the signature.

Similar arguments hold for the addition of ontological elements to the signature. Such elements are relevant to the domain conceptualized by the ontology at hand, since they are added to the signature, even if they do not (yet) appear in the axiomatic part. Thus, a language-level addition need not be coupled with a KB-level addition.

The identification of the exact side-effects of each operation in each level is irrelevant to this work and is omitted; the interested reader is referred to the standard ontology evolution literature (e.g., [9]) for a more detailed analysis of this issue.

2.3 Discussion on the Change Levels

As already mentioned, ontology evolution treats both language-level and KB-level operations in the same way. The analysis performed in the previous subsection implies that this approach may not be entirely correct from a methodological point of view, because it causes a mixture of effects upon both the axiomatic part of the ontology (KB-level) and the signature (language-level). The author argues that, even though both types of operations are useful, side-effects from one change level to the other should be avoided.

The argument can be stated more clearly with an example. Suppose that we attempt to develop an *ALC* ontology (see [1] for details on *ALC*), but later discover that we need more expressive power than the one provided by *ALC* for the particular domain. In that case, we are expected to switch to a new DL before adding any axiom types not supported by *ALC*. For example, if we want to add the axiom "$A \sqsubseteq B \sqcup \{x\}$" in the original ontology, we have to change the underlying DL first, then add the axiom.

If, instead, we attempted to add the new axiom directly, before changing (manually) the logic, that would not cause the introduction of the operator set-of ($\{\ldots\}$) into the underlying DL as a side-effect; no side-effect could cause a change in the underlying DL (logic-level change). On the contrary, the underlying ontology evolution system would not allow such a change (i.e., the addition of the axiom "$A \sqsubseteq B \sqcup \{x\}$" would be rejected as invalid).

What happens in this example is that a KB-level change is blocked (rejected) because it has a logic-level side-effect. This is considered intuitively adequate. But then, why should the addition of the axiom "$A \sqsubseteq B$" in an ontology whose signature does not contain B be allowed and cause the addition of B as a new concept (i.e., a KB-level change causing a language-level side-effect)?

Now consider a different case: suppose that the ontology engineer decides to switch logic by removing an operator (say $\sqcap$) from the DL. This, of course, should be made manually, as ontology evolution does not support logic-level changes. After such a change, much of the original ontology would be rendered invalid, as several axioms may use the removed operator. Nevertheless, we would expect the ontology engineer (rather than the ontology evolution system) to manually amend the axioms containing this operator so as to capture (as much as possible) the intended meaning of the axioms of the more expressive logic (the one containing $\sqcap$) using the axioms of the less expressive one (the one not containing $\sqcap$); this should be made before the removal of the operator $\sqcap$ from the logic.

On the contrary, we expect an ontology evolution algorithm to apply KB-level changes as side-effects in order to amend the axioms that are rendered invalid following the removal of a signature element (language-level change).

The conclusion from these examples is that there should exist clear boundaries between the various change levels disallowing the propagation of any side-effects from one level to the other. Should a change in one level cause changes in another level, it should be blocked or rejected until the knowledge engineer is given the chance to correct the problem(s) using change operations of the appropriate level.

This viewpoint is influenced by the viewpoint employed in standard logical formalisms. In belief change, only KB-level changes are considered: any changes that affect other levels, or that have side-effects in other levels, are rejected as non-valid. In fact, the operation "remove the predicate P from the language" would sound equally absurd to a logician as the operation "remove the operator $\sqcap$ from the DL" would sound to an ontology engineer.

The fact that belief change does not deal with language-level operations should not be viewed as a shortcoming of the field. If we confine each type of change to its own level only (by disallowing side-effects to other levels), then language-level operations become trivial to execute, because their language-level side-effects can be easily identified and resolved. Indeed, the removal of an element has no language-level side-effects, while the addition of an element could have, but only if the same name is already in use.

For example, if we are asked to add a class named P and there is already a property with that name, we should first remove the property before adding the class, as most formalisms (e.g., DLs) require the names used for classes, properties and individuals to be mutually disjoint. This side-effect would not exist in formalisms without this restriction, e.g., in RDF [14] or OWL Full [3]. In any case, such side-effects are trivial to identify, so belief change chose to ignore them. Of course, a language-level operation (in particular, a removal) could have a number of non-trivial KB-level side-effects, if such side-effects were allowed.

Another problem with language-level operations is that, unlike KB-level operations, it is not possible to formally describe a language-level operation using DL (or FOL) constructs. One of the consequences of this fact is that such operations render the recently proposed mapping of ontology evolution to belief change [7] unusable, since it is not possible to express a language-level change in the terminology used in belief change (even if it was, it wouldn't be of much use, as belief change does not provide any tools to handle such operations). A side-effect of this fact is that many formal approaches to ontology evolution (e.g., [10], [13], [17], [18]) do not consider language-level operations.

3. ALTERNATIVE REPRESENTATIONS

The previous section identified the need to keep operations affecting different levels separate and disallow side-effects from one level to affect the other. Even though such a rule is useful for the formal analysis of change operations, many existing methods do violate it.

In this section, we address this problem by describing two alternative techniques for representing ontologies. These representations allow the encapsulation of signature information into the axiomatic part of an ontology, which, in turn, confines both language-level and KB-level change operations (and side-effects) into the KB-level.

This way, we only need to consider KB-level operations which are well-studied and supported by both ontology evolution and belief revision, while still being able to perform changes (and side-effects) that would normally be classified as language-level ones. This allows us to enjoy the best of both worlds, since all useful operations and their side-effects can be addressed on the same level.

Applying these representation to ontologies has other advantages as well. First, it allows belief change techniques to be used to handle language-level operations; second, it makes the embedding of ontology evolution techniques into belief change methodologies (and vice-versa) possible; third, it allows a homogeneous treatment of all interesting operations; and, fourth, it allows methodologies originally designed to handle KB-level operations only to be used for language-level operations as well.

These representations should mainly take into account two important characteristics of signatures: first, there could be elements that are relevant to the ontological conceptualization (so they should appear in the signature in the standard approach), but for which no useful information is known (yet), so they don't appear in any of the "standard" DL axioms; second, the introduction of language-level assertions in the KB-level would inevitably introduce some non-standard KB-level information, whose semantics should be taken into account by the inference mechanism of the logic at hand.

Not surprisingly, the proposed alternative representations are not without problems of their own, discussed in the respective subsections. Such drawbacks are inherent in this approach, since this is actually an effort to model (represent) two intrinsically different types (levels) of information in the same representational level. Nevertheless, the proposed representations constitute interesting possible solutions to the problems described in the previous section because they allow the collapse of two representation levels into one. Both alternatives below will be described for DLs; however, they can be straightforwardly used for other logics as well, both in the logical and ontological setting.

3.1 First Alternative

This alternative originally appeared in earlier works by the author [5], [6], [7] in order to allow the representation of language-level ontology evolution operations using KB-level constructs. This was necessary to the end of being able to define the problem of ontology evolution in terms of the related field of belief change, which was one of the main objectives of the aforementioned works. Without the use of this alternative representation, only the part of ontology evolution dealing with KB-level changes can be described in terms of belief change.

Under this approach, the ontological signature is assumed static and the same for all ontologies; in particular, it is assumed that an ontological signature contains all possible element names (i.e., all strings of finite length). This deprives the signature from its original purpose of determining relevance of element names to the domain and raises the issue of how can one determine relevant and non-relevant element names.

There are two ways to resolve this problem. The first is to assume that there is no issue of relevance. All elements are, in principle, relevant to the domain of discourse, even though, for some of them, no information is known (yet), so they don't appear in any axiom. This approach was termed the *Open Vocabulary Assumption* (OVA) in [5]. Obviously, OVA causes the loss of all signature information and renders all language-level operators invalid, so it is not adequate for the purposes of this paper.

The second approach incorporates a new unary connective in the underlying DL to denote relevance; this connective is called the *Existence Assertion Connective* and is denoted by %. The semantics of % is that the axiom "%A" should be implied by the ontology if and only if the element A is relevant to the conceptualization of the ontology (i.e., it would have been part of the signature, if the standard approach was used). Using this connective, we can determine whether an element is relevant to the ontology or not, leading to what was termed the *Closed Vocabulary Assumption* (CVA) [5].

Of course, the standard DL inference mechanism should be amended in order to incorporate the semantics of the new connective. In [5] the proper amendments were described, which eventually boil down to two conditions: the first guarantees that whenever an element A appears in a "standard" DL axiom, then this DL axiom implies the "relevance" of the element (i.e., %A) but not the relevance of any elements not appearing in the axiom (e.g., %B); the second guarantees that axioms of the form "%A" do not imply any "useful" KB-level information, in the sense that no non-tautological "standard" axiom can be implied by any set of assertions of the form %A.

It is clear that the % connective "downgrades" language-level assertions into KB-level assertions, thus making possible the representation of what should be language-level change operations (and statements) using KB-level change operations (and statements). For example, the addition of an element A is now expressed as the addition of the axiom %A. The semantics of the inference relation dictate what the side-effects of such operations should be. For example, the removal of %A implies the removal of all axioms that include A (otherwise %A would re-emerge as an implication of such an axiom, due to the first amendment of the inference relation described above).

This fact implies that it is easy to adapt some standard belief change or ontology evolution algorithms so as to deal with language-level operations; all we have to do is replace the standard inference relation of the underlying logic/DL with the modified one. Of course, this technique may work only for the algorithms that are not tied to any particular logic/DL (and thus a particular inference relation).

The major disadvantage of this method is that it requires the addition of a non-standard connective in the logic, thus rendering standard inference algorithms non-sound for inferences that involve "fresh" elements, as well as non-complete for inferences

that involve the existence assertion connective. On the other hand, it is relatively easy to implement and it is applicable to any logic.

It is possible, even though not necessary, to refine the connective % so as to indicate whether an element is a class, role or individual (in effect introducing three different existence assertion connectives). Unfortunately, this refinement introduces an additional (and unnecessary) complexity in the approach so it will not be considered here. For a more detailed discussion on this refinement, as well as on the other issues raised in this subsection, see [5].

3.2 Second Alternative

This alternative maps DL information into FOL formulas, but, instead of using the standard mapping [2], it employs a twist in the way signature elements are viewed, resulting to a different mapping. This non-standard mapping has the advantage that it encapsulates the signature structure and allows it to be part of the resulting FOL KB. The final result is similar to the previous alternative: language-level assertions (change operations) can be expressed using KB-level assertions (change operations).

In order to implement this alternative, a FOL is defined whose language contains one predicate name for each connective appearing in the DL and one function name for each operator appearing in the DL. It also contains an infinite number of individual names (constants), which will be used to represent all possible element names that may appear in the ontological signature. To cover all cases, any finite-length string will be assumed to be a constant in said FOL (except, of course, from the symbols reserved for functions and predicates). In addition, the unary predicates Class(.), Property(.) and Instance(.) are included in order to capture language-level assertions, i.e., that a respective element name (a FOL constant in this representation) is a class, property or instance respectively in the DL ontology.

The mapping of a DL axiom into this FOL is made by rewriting the axiom using prefix (Polish) notation and then replacing each connective and operator with its respective predicate or function in the defined FOL. For example the axiom: "$\forall R.A \sqcap B \sqsubseteq C \sqcap A$" would be mapped into the FOL formula: "$Con_{\sqsubseteq}(Oper_{\sqcap}(Oper_{\forall}(R,A),B),Oper_{\sqcap}(C,A))$", where $Con_{\sqsubseteq}(.,.)$ is the binary predicate attached to the DL connective $\sqsubseteq$ and $Oper_{\sqcap}(.,.)$, $Oper_{\forall}(.,.)$ are the binary functions attached to the DL operators $\sqcap$, $\forall$ respectively. Language-level assertions are simpler to capture: Class(A), Property(A), Individual(A) imply that A is a class, property, individual respectively.

The mapping of axioms and signature assertions to FOL ground facts in the above manner is not enough, because the semantics of the connectives and operators are not carried over. To achieve this, the FOL KB should be coupled with a number of integrity constraints guaranteeing the intuitively expected behavior of the various FOL predicates and functions. For example, to guarantee the transitive semantics of the $Con_{\sqsubseteq}$ predicate, we need the constraint: "$\forall x,y,z\ Con_{\sqsubseteq}(x,y) \wedge Con_{\sqsubseteq}(y,z) \rightarrow Con_{\sqsubseteq}(x,y,z)$".

Similar constraints must be defined for the special predicates Class, Property and Instance as well; the general idea is the same as the one employed in order to amend the inference relation of the previous alternative. Unfortunately, the constraints in this case cannot be simplified by dropping the three predicates and keeping just one as was done in the previous subsection; such a change would not allow the detection of the invalidity of the statement "$Con_{\sqsubseteq}(Oper_{\forall}(A,A),A)$, as it would not be possible to determine that A in this statement is used both as a class and as a role.

It is clear by the above analysis that, for very expressive DLs, the task of defining all the necessary integrity constraints is very difficult; therefore, the difficulties involved in applying this method are depending on the logic's expressiveness (unlike the first alternative). This constitutes the most important drawback of this alternative, and makes it more adequate for less expressive logical formalisms.

The role of "downgrading" the language-level assertions into KB-level ones (undertaken by the % connective in the previous approach) is now performed by the three special predicates Class, Property, Instance. The same general comments on how this allows language-level changes and how existing (belief change or ontology evolution) algorithms could be used to address such changes apply here.

4. EPILOGUE

In this paper, three different representation levels were introduced (logic, language and KB) and an important distinction between changes affecting each level was introduced. This discussion is particularly relevant for the signature (language-level changes) and the axiomatic part of an ontology (KB-level changes); arguments were provided in favor of the discrimination of the two change types, as well as against allowing side-effects caused by a change to affect other levels.

Moreover, two alternative representation techniques were introduced that allow the collapse of the two lower levels (language and KB) into one (KB). These methodologies allow us to execute both language-level and KB-level changes at the same level (KB) and avoid the problem of side-effects caused from one level to affect another. In addition, these approaches facilitate the smoother integration of ontology evolution (dealing with language-level and KB-level changes) and belief change (dealing with KB-level changes only) approaches [7] and allow us to use methods originally designed to handle KB-level changes for language-level changes as well.

Even though these alternative representations suffer from various deficiencies, they could prove useful when the aforementioned collapse of the two levels into one is necessary. The deficiencies of the proposed alternatives show the inherent difficulty of this task and serve as an additional argument in favor of the proposed definition of representation and change levels.

The discrimination of the three representation levels is a known issue in the literature, but, to the best of the author's knowledge, the explicit classification of the various types of changes in three levels based on the representation level they affect was never considered before, except only superficially by earlier works of the author [5], [6], [7], as well as in [12], where a similar problem (variable forgetting) was addressed in the context of Propositional Logic.

5. ACKNOWLEDGMENTS

Significant fragments of this work have benefited from discussions with Carlo Meghini, Dimitris Plexousakis, Vassilis Christophides and Yannis Tzitzikas. This work was carried out during the author's tenure of an ERCIM "Alain Bensoussan" Fellowship Program.

6. REFERENCES

[1] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P. (eds). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2002.

[2] Borgida, A. On the Relative Expressiveness of Description Logics and Predicate Logics. *Artificial Intelligence, 82*, 1996, 353-367.

[3] Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuiness, D., Patel-Schneider, P., and Stein, L. A. OWL Web Ontology Language Reference. W3C Recommendation, 2004. Available at: http://www.w3.org/TR/owl-ref

[4] Enderton, H. B. *A Mathematical Introduction to Logic*. Academic Press, New York, 1972.

[5] Flouris, G. *On Belief Change and Ontology Evolution*. Ph.D. Thesis, Department of Computer Science, University of Crete, 2006.

[6] Flouris, G., Plexousakis, D., and Antoniou, G. Generalizing the AGM Postulates: Preliminary Results and Applications. In *Proceedings of the 10th International Workshop on Non-Monotonic Reasoning (NMR-04)*, 2004, 171-179.

[7] Flouris, G., Plexousakis, D., and Antoniou, G. Evolving Ontology Evolution. In *Proceedings of the 32nd International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM-06)*, Invited Talk, 2006.

[8] Gärdenfors, P. Belief Revision: An Introduction. In Gärdenfors, P. (ed). *Belief Revision*, Cambridge University Press, 1992, 1-20.

[9] Haase, P., and Sure, Y. D3.1.1.b State of the Art on Ontology Evolution. SEKT Deliverable, 2004. Available at: http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/SEKT-D3.1.1.b.pdf

[10] Halaschek-Wiener, C., and Katz, Y. Belief Base Revision For Expressive Description Logics. In *Proceedings of OWL: Experiences and Directions 2006 (OWLED-06)*, 2006.

[11] Kalfoglou, Y., and Schorlemmer, M. Ontology Mapping: the State of the Art. *Knowledge Engineering Review (KER), 18, 1*, 2003, 1-31.

[12] Lang, J., Liberatore, P., and Marquis, P. Propositional Independence: Formula-Variable Independence and Forgetting. *Journal of Artificial Intelligence Research (JAIR), 18*, 2003, 391-443.

[13] Meyer, T., Lee, K., and Booth, R. Knowledge Integration for Description Logics. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05)*, 2005.

[14] Miller, E., Swick, R., and Brickley, D. Resource Description Framework (RDF) / W3C Semantic Web Activity. 2006. Available at: http://www.w3.org/RDF

[15] Newell, A. The Knowledge Level. *Artificial Intelligence, 18, 1*, 1982.

[16] Plessers, P., de Troyer, O., and Casteleyn, S. Event-based Modeling of Evolution for Semantic-driven Systems. In *Proceedings of the 17th Conference on Advanced Information Systems Engineering (CAiSE-05)*, 2005, 63-76.

[17] Qi, G., Liu, W., and Bell, D. A. A Revision-Based Approach for Handling Inconsistency in Description Logics. In *Proceedings of the 11th International Workshop on Non-Monotonic Reasoning (NMR-06)*, 2006.

[18] Qi, G., Liu, W., and Bell, D. A. Knowledge Base Revision in Description Logics. In *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA-06)*, 2006.

[19] Stojanovic, L., Maedche, A., Motik, B., and Stojanovic, N. User-driven Ontology Evolution Management. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW-02)*, 2002, 285-300.

[20] Stuckenschmidt, H., and Klein, M. Integrity and Change in Modular Ontologies. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, 2003.

Workshop on

Web Services Interactions, Quality and SLAs (WS-IQS 2007)

26th of April 2007,
Poznań, Poland

Workshop Program Committee

Sami Bhiri, National University of Ireland in Galway, Ireland
Irene Celino, CEFRIEL - Politecnico of Milano, Italy
Dario Cerizza, CEFRIEL - Politecnico of Milano, Italy
Monika Kaczmarek, Poznan University of Economics, Poland
Natallia Kokash, University of Trento, Italy
Andre Ludwig, University of Leipzig, Germany
Xuan Thang Nguyen, Swinbourne University of Technology, Australia
Dominik Zyskowski, Poznan University of Economics, Poland

Quality of Protection Determination for Web Services *

Artsiom Yautsiukhin
University of Trento
evtiukhi@dit.unitn.it

ABSTRACT

Security is a very important aspect for Web Service technology. There are a large number of works devoted to security of Web Service transactions. However, we argue that security must be guaranteed for data processing (after transmission) as well. These requirements must be negotiated with a client and inserted into the agreement between a client and a contractor. The problem is that a client and a contractor have different views on how these requirements should look like. We propose a methodology which binds these views and describes a process for selection the security configuration that helps to achieve negotiated level of protection.

1. INTRODUCTION

Web Services is a rapidly emerging technology which has been developed to simplify business-to-business integration. It has a great potential to facilitate IT business outsourcing, when processing of an IT work package is delegated to an external organization. One of the important issues for Web Services is to shift relationships between involved parties to contractual ones. The first step in this direction is an unambiguous and clear definition of a *Service Level Agreement (SLA)* between a client and a contractor reflecting desired *Quality of Service* (QoS) (e.g. performance, maintenance). For this purpose XML-based specifications WS-Agreement [1] and SLAng [12] providing templates to describe QoS were proposed.

We would like to focus reader's attention on security requirements which should be inserted in the agreement. According to established standards (WS-Security [2], WS-Security Policy [6]), security requirements for Web Services are specified as policies which must be fulfilled in order to get access to the service. WS security standards do not mention data protection after transmission. The data may be corrupted during processing on contractor's server because of careless security management (e.g. data can be stored in a server without a properly configured antivirus). We argue that SLA must be extended with the section of an agreement that contains security requirements, which is called *Protection Level Agreement* (PLA). Similarly to QoS, we define *Quality of Protection* (QoP) as a set of security requirements a PLA guarantees. For more details we refer the reader to our previous work [10].

In this paper we provide a methodology for the aggregation of security requirements. It helps to select the most suitable security configuration according to a contractor's business process and different levels of trust between involved partners. The proposed methodology captures and binds security requirements useful for contractors with ones understandable by clients. Supported by a reasoning algorithm the methodology will be able to evaluate possible security system configurations. It will allow the contractor easily recalculate his QoP if a partner or his trust level has been changed or small system reconfigurations made.

The paper is organized as follows. In Section 2 we define a problem which emerges because a client and a contractor have different viewpoints on PLA. In Section 3 we propose our methodology where we: provide a strategy for QoP hypergraph contraction (Subsections 3.1), define a propagation function for the hypergraph (Subsection 3.2), decompose security services and link them with QoP hypergraph (Subsection 3.3) and briefly discuss how the algorithm for root QoP calculation should be implemented (Subsection 3.4). In the last section conclusions and future work are outlined.

2. PROBLEM

The crucial point in PLA negotiation is the identification of metrics which describe the level of protection. We have found useful to divide all metrics into two types:

- *Internal metrics* describe security qualities used by a contractor to achieve a high level of security.
- *External metrics* are negotiated with the client to show that her security requirements are addressed.

Some examples of internal metrics are: time between updates, length of passwords, percentage of compliance with a standard [7]. Possible examples of external metrics are number of successful attacks on client's data confidentiality [4] and mean time to intrusion affecting client's data [13].

*This work was partly supported by the project EU-IST-IP-SERENITY, contract N 27587

The main problem is that internal metrics are not informative enough for a client because they do not state explicitly how her assets will be affected by breaches in contractor's security system. On the other hand, external metrics do not tell the contractor how he should configure his system to achieve the metrics. The contractor must map the external metrics negotiated with client (PLA) to a functional security SLA to receive concrete requirements for security system configuration. In a sequel we will call the functional security requirements as *Qualities of Security Service (QoSS)*.

3. BINDING METHODOLOGY

We propose a methodology which helps a contractor to determine a QoSS satisfying the PLA negotiated with a client. In our methodology we use directed hypergraphs to capture structure of contractor's business process which determines how security requirements are distributed among its activities. A directed hypergraph is a generalization of directed graph where edges (or hyperedges) start from a *set* of nodes (source nodes) end end at a single node (target node) [3].

In our methodology we assume that a contractor and a client have negotiated a PLA using external metrics. We also assume that a contractor has a business process (BP) written in a hierarchical way. In other words, a provider defines a high level (abstract) BP (BP_h) where all activities are connected with one structural pattern (i.e. "sequence", "switch", "while", "flow"). Then for each non-atomic activity A_i a BP (BP_{A_i}) is determined. The decomposition continues until atomic activities are reached.

3.1 Phase 1. Build a QoP hypergraph

In the first phase of our methodology a contractor breaks down the requirements stated in the PLA into more fine-grained ones according to the business process and represents them as a hypergraph.

Security requirements are identified for each activity of BP_h and connected with a top QoP node (PLA). We show this as a hyperedge from the requirements for the activities to the top QoP node for "flow", "sequence", "switch" and "while" patterns. Then we repeat the process for each activity and its sub-process. If design alternatives for the decomposition exist they are represented as several hyperedges.

Different partners to whom some services (parts of the BP) are outsourced have various level of trust. This fact also impacts identification of target metric values. A contractor may trust one partner that the defined metrics for the activity will be achieved and not trust another one. We use the following strategy to take this fact into consideration: if the contractor does not trust a partner that some QoP requirements will be achieved he should increase the estimated bound of the external metrics. Now the contractor may trust more the partner since the requirements is more likely to be met. In the hypergraph a partner is represented as an extra node between the target node and source ones or simply as a node connected with the target node if the subprocess for the outsourced activity is not known. If there are several partners who fulfill the same activity we use one hyperedge, when several alternative partners are connected to the target node with several distinct hyperedges. The algorithm for the process is shown in Figure 1. It takes a set of business processes S_{BP} and a set of activities A and returns a QoP hypergraph $H = (N, E)$ where N is a set of nodes and E is a set of hyperedges.

EXAMPLE 1. *Let us consider the following e-banking scenario. A holding company (customer) outsources task of providing a loan to one of its subsidiaries (contractor). The procedure is implemented using Web services. The subsidiary specifies a business process shown on the left side in Figure 2. The contract between the partners states that no more than 10 frauds may occur per one year of providing the service. To determine if it can meet this requirement the subsidiary first creates a QoP hypergraph as it is shown in Figure 2. The defined process is not finite because there are several design alternatives. First, the subsidiary has to select the credit bureau it will invoke to receive trustworthiness rating of a client. Second, the subsidiary may prepare a loan for all clients in the same way, or to prepare a loan for ordinary clients when the procedure for VIP persons is provided by a special department. Note, that the alternatives are shown in the figure as separate hyperedges leading to the same target node. The process of VIP department is known because it is under the subsidiary's control while credit bureaus are black boxes for the subsidiary.*

3.2 Phase 2. Propagation function assignment

Now we define semantics for QoP hypergraph. For each hyperedge a weight that shows contribution of a source node to the target one is assigned. Weights of edges connecting partners with a target node specify the level of trust between the delegator and the delegatee. Since in our case each source QoP node contributes differently to the target QoP one we use intermediate nodes between source and target nodes. The weights are assigned to the edges which connect source and intermediate nodes and the weights for the edge between the latter and target nodes are neutral (e.g., 1). We do not depict the nodes in the figures to avoid unnecessary complexity.

For all nodes we assign a tuple $\langle M_{QoP}, f_{QoP} \rangle$ where M_{QoP} is a vector of metric values which *can* be achieved if a specific QoSS is applied; f_{QoP} is a propagation function which computes a set of metric values M_{QoP} of the target node taking source nodes' M_{QoP}s and corresponding weights as arguments: $f_{QoP} : 2^W \times 2^{M_{QoP}} \mapsto M_{QoP}$. This function is different for the four basic structural patterns but it is defined in the same way for the same pattern. The functions depend on type of requirements and we are going to specify them in the future work. If an activity is outsourced the meaning of the function is how security requirements are changed according to trust level of the partner. These functions are determined by security staff using their experience, events history and modern trends.

3.3 Phase 3. Security services identification and decomposition

In this phase a contractor identifies security services which he has to provide to achieve requirements stated in the PLA. First of all, security services which can be implemented or which are already in place are determined. For each security service a set of security service parameters ($QoSS$) is

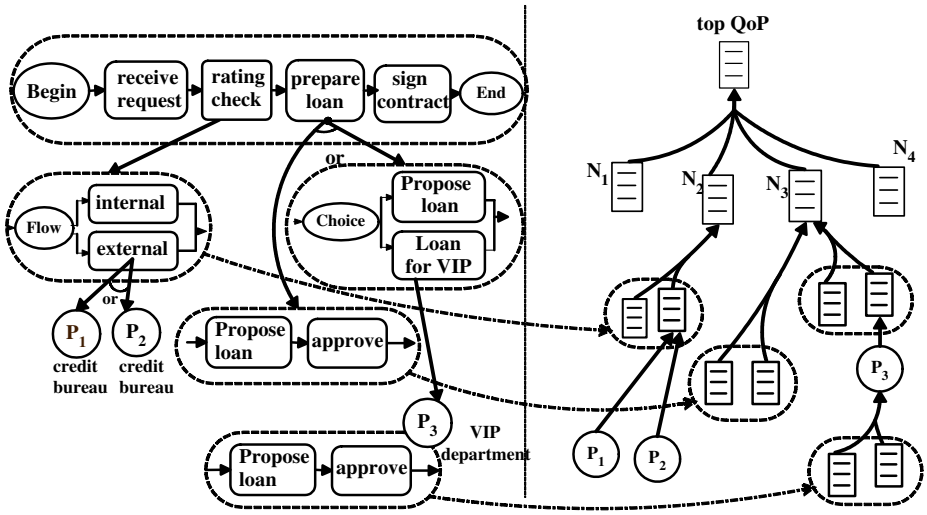

Figure 2: Building QoP hypergraph

Build_QoP_Hypergraph
input S_{BP}, A
  Add a new node QoP to N;
  $New\_Branch(S_{BP}, A, S_{BP}[1], QoP)$;
    `//Start with` BP_h $(S_{BP}[1])$
output N, E

New_Branch
input $S_{BP}, A, BP, TargetQoP$
  for all activities $A[j]$ **in** BP
    Add a new node QoP to N;
    Add node QoP to $SourceQoP$;
      `//set` $SourceQoP$ `is a tail of an edge`
    if the activity $A[j]$ is delegated **then**
      for all alternative sets of partners P_{alt} for $A[j]$
        `//for all edges connecting a set of`
        `//partners` P_{alt} `and target activity` $A[j]$
        for all partners p from set P_{alt}
          Add a new node QoP_1 to N
          Add node QoP_1 to $SourcePartner$
        `//set` $SourcePartners$ `is a tail of an edge`
        `//connecting a set of partners and` $A[j]$
          for all alternative BPs $S_{BP}[k]$ of p for $A[j]$
        `//`p `may fulfill` $A[j]$ `in different ways`
            $New\_Branch(S_{BP}, A, S_{BP}[k], QoP_1)$
        end
        Add an edge from $SourcePartner$ to QoP_1 in E
      end
    else
      for all refining BPs $S_{BP}[k]$ for activity $A[j]$)
        $New\_Branch(S_{BP}, A, S_{BP}[k], QoP)$
      end
    end
  Add an edge from $SourceQoP$ to $TargetQoP$ in E
  end
output N, E

Figure 1: QoP hypergraph building algorithm

determined. These parameters are internal security metrics of the service. Each compound service is decomposed in a similar way as it is shown in the first phase, so at the end we have a set of disjoint QoSS hypergraphs. A propagation function is assigned to each QoSS node which denotes how source security services contribute to the target one.

The contractor links potential security services with leaf QoP nodes which can be achieved if the countermeasures are installed (Figure 3). These links show if the countermeasures help to satisfy a requirement ("+" mark) or deny it ("-" mark). For leaf QoP nodes we assign a propagation functions similar to the one for other QoP nodes. For those

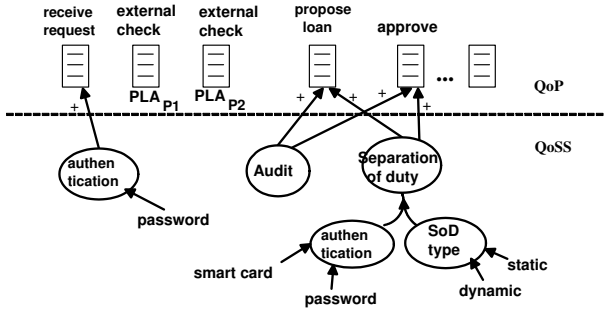

Figure 3: QoSS contribution.

leaf nodes which are delegated to other partners metric values can be taken from the corresponding PLAs. In case all tasks are outsourced (the contractor is a Web Services orchestrator) the methodology will choose those partners with which the overall process has the best protection level.

EXAMPLE 2. *In our example the security staff of the subsidiary have defined the the following security controls to reduce number of frauds: authentication of the client, audit of employees activity and separation of duty (to avoid approval*

of the loan by the same person that proposes it). Note, that for "external rating check" activity metric values are taken from PLAs of the credit bureaus.

3.4 Phase 4. A reasoning algorithm.

We apply a reasoning algorithm for testing different security configurations and determination of the best one. The contractor chooses a set of security services he is going to provide and determines security parameters of the leaf QoSS nodes. Using QoSS propagation function top security services are derived. Then the metrics for each leaf QoP node are calculated or determined according to PLAs for outsourced services. Now we have a classical problem of finding the shortest hyperpath in a hypergraph for which efficient algorithms have been proposed (e.g., [3]). Note, that these algorithms can be used only for those metrics for which QoP propagation functions are superior/inferior (e.g. number of attacks per execution). In the future work we are going to adopt the algorithms for other metrics (e.g., number of attacks per month). Finally, we receive the best value of the top QoP node. If the calculated protection level is less than the one agreed in the PLA with a client then another security configuration is tested. The process may be automated (to avoid manual correction of security parameters) but this direction requires further investigation such as definition of satisfaction function and security parameter correction mechanism.

4. RELATED WORK

There are a few papers which tackle the issue of security requirements in business outsourcing. One of the first papers discussing security SLA in a large enterprise is [9]. The main idea is to check compliance the system with fifteen security domains split into best practices. For each best practice the security service level is determined and added to the SLA (yet it does not consider outsourcing). Casola et. al. [5] extend the security decomposition to compare two SLAs or to find a security SLA which is the closest to the desired one. A similar idea was applied to evaluation of Web Service security by Wang and Ray [14]. Karjoth et. al. [11] claimed that security requirements must be reflected in the contract. *Trusted Virtual Domains (TVDs)* [8] are intended to connect a number of remote trustable virtual processing environments in one secure network. Security operational policy (accord of PLA/SLA), which is obligatory for every environment, are used. This technology can be applied to client-contractor interaction when one side (most likely, a contractor) allows another one to use its TVD.

5. CONCLUSION AND FUTURE WORK

In this work we have described the methodology which helps a contractor to determine the security system configuration that fulfills the requirements negotiated with a client. The methodology binds internal security requirements useful for a contractor with the external ones understandable by a client. It also allows a contractor easily recalculate security level if changes in a system configuration occur.

In future work we are going to define a propagation function for three basic business process constructs. We are also going to implement the algorithm adopted for chosen functions and test effectiveness and correctness of our approach.

6. ACKNOWLEDGEMENTS

I would like to thank my advisor Fabio Massacci for invaluable help in conducting this research.

7. REFERENCES

[1] A. Andrieux et. al. *Web Services Agreement Specification (WS-Agreement)*. Global Grid Forum, 2 edition, August 2004.

[2] B. Atkinson et. al. *Web Services Security*. Microsoft, IBM, VeriSign, 1.0 edition, April 2002.

[3] G. Ausiello, G. F. Italiano, and U. Nanni. Optimal traversal of directed hypergraphs. Technical Report TR-92-073, Pisa University and Monreal University, Berkeley, CA, 1992.

[4] S. A. Butler. Security attribute evaluation method. Technical Report CMU-CS-03-132, Carnegie Mellon University, May 2003.

[5] V. Casola, A. Mazzeo, N. Mazzocca, and M. Rak. A SLA evaluation methodology in Service Oriented Architectures. In *Proceedings of the 1st Workshop on Quality of Protection.*, Milan, Italy, 2005. Springer-Verlag.

[6] G. Della-Libera et. al. *Web Services Security Policy Language*. IBM and Microsoft and RSA Security and VeriSign, 2005.

[7] J. Eloff and M. Eloff. Information Security Management - A New Paradigm. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists*, pages 130 – 136, 2003.

[8] J. L. Griffin, T. Jaeger, R. Perez, R. Sailer, L. van Doorn, and R. Cáceres. Trusted virtual domains: Toward secure distributed services. In *Proceedings of the 1st Workshop on Hot Topics in System Dependability*, Yokohama, Japan, June 2005.

[9] R. Henning. Security service level agreements: quantifiable security for the enterprise? In *Proceedings of the 1999 Workshop on New security paradigms*, pages 54–60. ACM Press, 2000.

[10] Y. Karabulut, F. Kerschbaum, P. Robinson, F. Massacci, and A. Yautsiukhin. Security and trust in it business outsourcing: a manifesto. In *Proceedings of the 2nd International Workshop on Security and Trust Management. To appear.* Electronic Notes in Theoretical Computer Science, 2006.

[11] G. Karjoth et. al. Service-oriented assurance comprehensive security by explicit assurances. In *Proceedings of the 1st Workshop on Quality of Protection.*, Milan, Italy, September 2005. Springer-Verlag.

[12] D. D. Lamanna, J. Skene, and W. Emmerich. SLAng: A Language for Defining Service Level Agreements. In *Proceedings of the The Ninth IEEE Workshop on Future Trends of Distributed Computing Systems*, pages 100–18. IEEE Computer Society Press, 2003.

[13] R. Ortalo, Y. Deswarte, and M. Kaaniche. Experimenting with quantitative evaluation tools for monitoring operational security. *IEEE Transactions on Software Engineering*, 25(5):633–650, 1999.

[14] Y. Wang and P. K. Ray. Evaluation methodology for the security of e-finance systems. In *Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service* . IEEE Press, 2005.