A Socio-Technical Approach for Topic Community Member Selection

Aldo de Moor[1] and Anjo Anjewierden[2]

[1] CommunitySense, Tilburg, The Netherlands
[2] University of Twente, The Netherlands[3]

1. Introduction

Wicked problems and social complexity abound in our globalizing, ever more complex society [6]. Wicked problems, such as many socio-economic and environmental issues, cannot be solved in traditional ways, as no perfect solution can be found. Also, the understanding of the problem evolves as the solution is being worked on, but no clear agreement on what the real problem is can be reached. The only way to seriously address these problems is by examining a wide range of possible solutions, argumentations, and viewpoints by as many stakeholders as possible [13, 6]. Classical organizations, like governments and official scientific bodies, are no longer capable of representing these interests on their own. New forms of agile social structures are needed, covering a wide spectrum of public interests instead of limited national or organizational interests.

Communities, as cornerstones of civil society, play a key role in representing the public interest. They do for the public good what markets do on behalf of aggregated private interests [4]. Communities have discovered the bottom-up use of cyberspace, a space of distributed power, as a way to transform society [16], making them important catalysts of societal

[3] A first version of this paper was written by the first author while working at STARLab, Vrije Universiteit Brussel, by the second author while working at the Dept. of Human Computer Studies, University of Amsterdam.

Presented at the 3rd International Conference on Communities and Technologies, Michigan State University, East Lansing, Michigan, June 28-30, 2007

Published in Steinfield, Pentland, Ackerman, and Contractor (eds.), Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference, Michigan State University, 2007, London: Springer, 225-244.

change. Communities can revitalize public discourse, as Habermas, for instance, argues in his theory of communicative action [8]. A good example are the conversational practices in numerous and often influential blog communities [9].

To address the multitude of very complex and interconnected political, socio-economic and environmental issues, *topic communities* of experts and stakeholders play an increasingly important role. These topic communities are often created ad hoc and urgently, requiring their members to have a unique mix of experience and expertise. They can be characterized as being interdisciplinary, informal, and of global scope, having an urgent need for action and often very limited resources. This means that it is crucial that the most relevant experts and stakeholders are discovered at the lowest possible cost. However, the generation of such communities is far from trivial. Existing scientific and political frameworks do not suffice to provide the right experts and stakeholders.

A necessary condition for building high quality collaborative communities is selecting the best possible candidates, who should be leaders in expertise and experience, and have shown a commitment to public debate. A good place to look for such people is the global *blogosphere*. In this paper, we present a socio-technical approach for topic community member selection. It is based on the analysis of large corpora of blog posts to identify combinations of topics and bloggers likely to be relevant to the goals of the topic community. The technical module of the approach is based on the community layer of the tOKo tool for text analysis [2]. The social aspect consists of a sequence of steps of human interpretation of the subsequent blog analysis results. The resulting socio-technical process forms a "pragmatic funnel" leading to a relevant set of candidate topic community members.

In Sect. 2, we introduce our socio-technical approach for topic community member selection. In Sect. 3, we focus on the technical machinery of the blog analysis. In Sect. 4, we illustrate our approach with a hypothetical, but realistic case on an urgent worldwide problem: ice cap melting. We end the paper with a discussion and conclusions.

2. A Socio-Technical Approach for Topic Community Member Selection

Communities can be important instruments for social knowledge creation and transformation, able to bridge communication gaps of many kinds [12]. We define a *topic community* as a community of interest consisting of

experts with expertise on and stakeholders with hands-on experience with the complex focal topic on which the community aims to build in-depth knowledge. Only such communities can effectively research causes, effects, and solutions of wicked problems, such as global warming, socio-economic development, and impacts of globalization on local communities, nations, and the world.

In creating such knowledge sharing communities, the focus should, at least initially, be on creating the conditions for the community to emerge, rather than on complex issues of responsibility, commitment, and reward [5]. The importance of this initial stage is recognized in many of the well-known community life cycle models, which distinguish explicit stages like planning, potential, forming, and committing [14, 11, 20]. Crucial in this stage of a knowledge community is that people face similar situations, problems, and interests without the benefit of a shared practice [22]. The question thus is: how to recognize which people face similar situations without them yet being involved together in knowledge sharing activities?

Most knowledge creation communities emerge in an organizational setting. This means that explicit structures, processes, and incentives are present for the community formation stage to take place. In such organizational settings, much experience has already been gathered on how to cultivate communities of practice, e.g. [23]. For example, pilot schemes have been proposed to introduce business weblogs in medium and large enterprises [15].

With topic communities, such a systematic, organization-supported approach does not work. These communities being low on resources and high on pressure, other approaches are needed to form their initial membership. In this paper, we focus on blog communities as a "substrate" on which many different topic communities could be grown. The millions of blogs in the blogosphere contain numerous, often very high quality weblog conversations: series of interlinked weblog posts and comments on a specific topic, usually not planned, but emerging spontaneously [9]. Also, multiple tools exist that gather, index, and classify these many resources[4]. Thus, weblogs provide a very important resource of knowledgeable people, experts and stakeholders, who might get involved in a particular topic community.

Still, despite this potential for mining the blogosphere, these tools provide only crude results, since basically they are based on simple term searches. This may be sufficient for casting a wide net when looking for particular terms. However, precision and recall from a point of view of identifying potentially *relevant* bloggers as candidates for topic communi-

[4] E.g. www.technorati.com, blogsearch.google.com, www.blogpulse.com

ties are totally insufficient. For example, say we are interested in forming a topic community on studying the dramatic problem of ice cap melting (we will work out this case in Sect. 4), then some important limitations of such an approach are that (1) the *concept* of ice cap melting has many different surface forms ("melting of the ice caps", "ice sheet melting", ...), which are not covered with a simple keyword search, (2) there are many related concepts which should also be taken into account when creating a topic community (e.g. it is very likely that blogs treating global warming as a likely cause of ice cap melting are also relevant) and (3) that many of the results, especially when using more common terms, provide a large set of unrelated posts, which need to be pruned in order not to pollute the data set.

It will be obvious that a much more sophisticated approach is necessary to identify a useful set of potential topic community members that is relatively precise. In this paper, we outline a \emph{socio-technical} approach for topic community member selection. It combines an advanced tool for community text analysis with human common sense interpretation. We show how, in a series of analysis and interpretation steps, a plausible set of potentially relevant topic community members can be detected. The approach consists of the following steps:

1. Preparation
   - Definition of topic community purpose
   - Selection of focal concept
   - Selection of document corpus
2. Concept Exploration
   - Exploration of concept network
   - Identification of key concepts to be used in ranking of documents
3. Document Exploration
   - Operationalization of key concepts in search terms
   - Pattern search of documents on key search terms
   - Ranking of documents
4. Community Exploration
   - Identification of document authors
   - Community analysis
     - Link analysis
     - Trend analysis
   - Ranking of document authors
   - Interpretation of ranking

The analysis techniques used in these steps are explained in the next section. In Sect. 4, we explain how these are applied in our socio-technical approach to potential member selection for a topic community on ice cap melting.

3. The Analysis of Digital Traces in Weblog Communities

The most common digital traces an ad hoc community leaves behind are documents: in particular on-line documents in the form of webpages. The tool we have used for the analysis of a corpus of documents by a community is called tOKo [3]. tOKo is an open source tool for text analysis, with support for ontology development and, using the extensions described in [2], exploring communities.

An example screenshot of tOKo is shown in Fig. 1. The figure shows the corpus of the case study as described in the next section. In general, tOKo supports both text analysis with the purpose of developing an ontology, *and* using an ontology to search in a corpus. The main mechanism to achieve some level of equivalence between concepts in the ontology and terms in the corpus, is to interpret a concept as a lexical lemma. For example, if the user defines **sea level rise** as a concept, tOKo automatically derives the lexical variations. When the user, thereafter, searches for the concept **sea level rise** it will also match the phrases **sea level rises**, **sea levels rise** and consider these as occurrences of the concept **sea level rise** in the text.

The figure provides an example of the browsing capabilities in tOKo. The second browser at the top lists lemmas in the corpus ordered by frequency, selecting a lemma (**climate**) shows the documents that contain the lemma in the top left browser. The rightmost browser at the top shows (possibly) compound terms related to **climate change** according to a co-occurrence metric.

In other words, the basic dimensions tOKo supports are *documents* and *terms* (and the derived notion of concepts based on lexical variants). For community analysis, however, these two dimensions are not sufficient. A framework for community analysis proposed in [2] distinguishes three additional dimensions:

Document	A self-contained publication by a member in the community. Examples of documents are a web page, email or weblog post.
Term	A meaningful term used by one or more members of the community. These terms occur in documents.
Person	A member of the community.
Link	A reference from one document to another document, and implicitly between the persons who authored the documents.
Time	The date, and possibly time, of publication of a document.

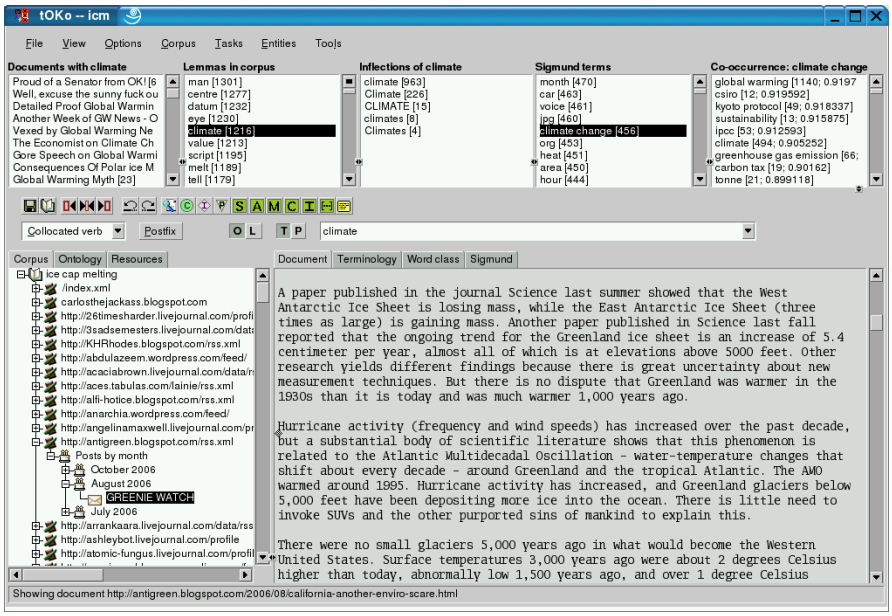

Fig. 1. tOKo Overview

The *person* dimension is critical for the present paper as we want to identify the level of expertise of a person as evidenced by the content of the digital traces. *Links* are an additional indicator of possible expertise or connectedness in a community. A document that is linked by several other members in the community is likely to be more relevant than a document that is not linked. Finally, the *time* dimension can be used to track changes of the topical focus and to identify various types of trends of term usage.

The framework thus focuses on communities that leave digital traces in the form of documents, and derives the other dimensions from metadata (person, time) and content (terms, links). Given a dataset represented along these dimensions a tOKo user can navigate through it by specifying one or more initial dimensions, fixating a particular dimension (e.g. focusing on a particular term, person, or time period). Navigating along multiple dimensions makes it possible for the researcher to obtain both an overall view (e.g. what are the most frequent terms used in the community) and more detailed views (e.g. term usage of a particular member over time).

4. Case: Selecting Potential Members of a Topic Community on Ice Cap Melting

tOKo is a powerful toolkit for the analysis of community text corpora. However, a hammer by itself does not do anything, it needs a work context in which it is used to create something. In this paper, we present the initial version of a necessarily much more complex approach that makes optimal use of the tOKo functionality for community analysis purposes. In this section, we operationalize the conceptual outline of our topic community member selection approach by demonstrating its possible use in a hypothetical, yet plausible scenario.

4.1 Scenario: Developing a Topic Community on Ice Cap Melting

Only a few years ago, much scepticism existed about whether global climate change is really happening. At the moment, scientific consensus is quickly being reached that the situation is even more dramatic than many had feared and that urgent measures are needed[5]. However, the causes and effects of climate change are still ill-understood, while feasible solutions to address these issues are even more behind the horizon.

Let us assume that the United Nations Environment Programme (UNEP) has been commissioned to - as quickly as possible - set up a set of task groups that focus on studying causes, effects, and possible solutions of a whole range of climate change-related issues. The task groups are to be broad in scope, made up of both scientific experts and opinion leaders from business, governments, NGOs and the general public. Their initial assignment is to come up with lists of practical proposals to be discussed during the next summit of government leaders. Over time, however, they should grow into true, collaborative topic communities, being able to evaluate a broad range of issues and policy recommendations and act as high-quality sounding boards for policy makers. Since the time for their setup is very limited, Jane, a senior UNEP official, accesses the *Topic Community Explorer* tool.

In the following, we suggest how - a more evolved version of - our socio-technical approach, could be implemented and used in practice.

[5] As unequivocally stated by the Intergovernmental Panel on Climate Change (IPCC), www.ipcc.ch

4.2 Applying the Socio-Technical Approach

1. **Preparation**
   In this stage, the generic topic community member selection approach is configured for the topic community at hand.

   - **Definition of topic community purpose**
     The purpose of the topic community is defined in terms of what the outcome of the analysis should be (theme, types of knowledge to be defined, types of experts and stakeholders to be invited, time frame, and so on). This knowledge is important for the human interpreters making the decisions about what are interesting concepts and terms when exploring the corpus, whether the focus is more on long-term general expertise, or current, very specialized expertise, audience of the topic community, and so on.

     Example: We are interested in those experts and stakeholders who prominently take part in the online debate and who are well versed in as many of the causes and effects of ice cap melting as possible, since general policy recommendations are desired.

   - **Selection of focal concept**
     The focal concept should be the key concept related to the purpose of the community. It should be wide enough to link to many of the causes, effects, and solutions of the problem for which the topic community is being established.

     Example: The focal concept in our case is simply **ice cap melting**, since the task group was being established especially for that purpose. Other topic communities may have more focal concepts, for each of which an analysis like proposed here can be performed.

   - **Selection of document corpus**
     A relevant online document corpus needs to be delineated and retrieved. Issues to be taken into account include: which search engine to access, what keywords to use, what time period to cover, etc.

     Example: Several sites provide a facility to search for blog posts given a certain query. We selected BlogPulse[6] as the engine of choice. BlogPulse has the advantage that it concentrates on blogs by persons, whereas Google blog search also includes pages from news sites.

     Blog posts were collected by querying BlogPulse for "ice cap melting" (749 hits) and then retrieving the HTML pages containing the

[6] www.blogpulse.com

posts. No further processing of the HTML pages was performed. That is, we did not attempt to remove the side matter that appears on most pages (e.g. blogrolls, tag clouds, advertisements, comments, etc.). One reason is that this is quite a complex operation to implement. A more principled reason is that this side matter often contains high-quality terms, reflecting the key interests of the poster (e.g. tag clouds). The resulting corpus of 19Mb was loaded into tOKo.

2. **Concept Exploration**
   In this stage, the concepts relevant for document and community exploration are identified.
   - **Exploration of concept network**
     The focal concept is surrounded by a network of concepts relevant to the purpose of the community. This analysis needs to be elaborate, since it is the main factor in determining the quality of the search results. tOKo includes many facilities to perform such an in-depth analysis.

     Example: The previous step has resulted in a large corpus of blog posts which contain words. Jane is interested in concepts, particularly concepts related to ``ice cap melting''. Sigmund [1] is a tool part of tOKo that can identify meaningful terms. A meaningful term is defined to be a noun phrase (without prepositional parts) or a proper noun (e.g. names of persons, locations, etc.). An important reason to use Sigmund is that most interesting terms, in more or less any domain, are compound.

     Sigmund returned 4211 compound terms with a frequency of least 10, many of which point to concepts the community is blogging about. Fig. 2 gives examples of terms automatically extracted, followed by their frequency.

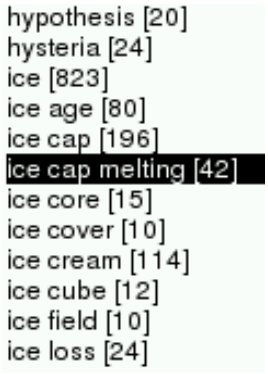

Fig. 2. Some significant terms found in the corpus (frequency between brackets

Co-occurrence is a measure that generally points to semantic similarity and Jane starts exploring Sigmund terms that co-occur with "ice cap melting". One of the terms is "global warming", and she starts exploring it, as she thinks it is relevant. From "global warming" she finds "climate change". Jumping from one concept to another is visually supported by a graphically represented network of terms with a high co-occurrence. A network of terms related to "climate change" is shown in Fig. 3.

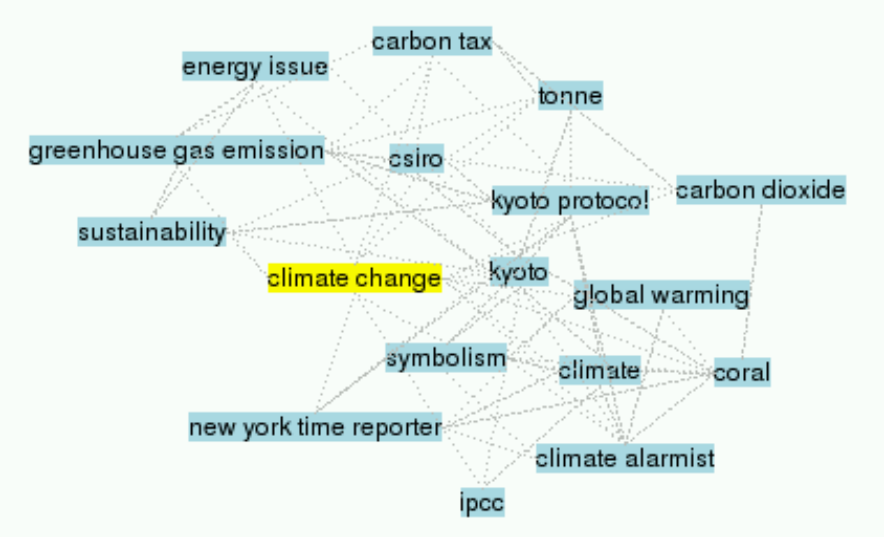

Fig. 3. Network of terms with a high co-occurrence to "climate change"

- **Identification of key concepts to be used in ranking of documents**
  Using the results of the previous concept exploration, the human interpreter decides which concepts are most relevant for the purposes of the topic community. It is important to realize that the concepts to choose from are derived from the corpus itself, so that they are already tailored to the interests of the blog communities underlying the topic community to be formed. It is thus ensured that this expertise is actually present, which is not the case when, say, an online thesaurus would be used instead.

  Example: When Jane sees a term in the network that might point to a key concept she adds it to the ontology (discussed further on). Later, these concepts are used to find the most interesting documents and the people related to the key concept.

3. **Document Exploration**
   Using the relevant conceptual network defined in the previous steps, the next step is to identify which documents best match these concepts.
   - **Operationalization of key concepts in search terms**
     The key concepts are those that not only co-occur or occur with high frequency, but those that are related. Language offers a multitude of mechanisms for expressing how the concepts are related and what the relation is. A trivial example is: ``a horse is an animal". Another one would be: "inflation causes price increases".

     Example: Jane, by chance, stumbled on ``ice cap melting which causes a sea level rise" in one of the posts. This phrase grammatically conforms to the following rule: **NP** *noise* **to cause** *noise* **NP** (NP is noun phrase).
   - **Pattern search of documents on key search terms**
     The document set is scanned for documents matching with the defined patterns.

     Example: Jane decides to use the tOKo pattern search facility in which the search rule can be specified. In the pattern search language[7], the above rule is `@ ... (cause) ... @`}, where `@` matches a Sigmund term which is effectively a noun phrase, `...` matches the noise and `(cause)` matches the verb **to cause**. Fig. 4 shows part of the results as a concordance table. The middle part matches the pattern and the left and right are the preceding and following text.
   - **Ranking of documents**
     Documents can be ranked on a wide range of criteria. A simple one is the number of matches with the set of search patterns.

     Example: Jane's main interest is in the document authors. Since her search result set is only relatively small, she decides not to put a threshold and accept all documents for further analysis.

4. **Community Exploration**
   The output of the previous steps is that several concepts that the interpreter thinks are relevant have been identified. In this stage, the authors of documents are analysed, both individually and their interrelationships.

[7] See www.toko-sigmund.org/pattern_search.html for details.

moderate latitudes. Neither the cause of the	ice ages nor the cause of the retreat	of the icy desert is known; the
i was just reading about the melting	ice caps cause a loss	of life in the antartic. these cuddly
hear environmentalists whine that the melting polar	ice caps will cause the oceans to rise	and flood the shores and wipe out
against countries that harbor terrorists, we harbor	CEO's whose companies cause far more misery	and suffering in the world than the
widespread in the future.Rising	global temperatures will cause the melting of glaciers	and ice caps, and cause early ice
atmosphere, caused by global warming. Although	global warming causes temperatures to rise	in the troposphere, the lower atmosphere,
way of life it creates for everyone	Global warming will cause great changes	to our atmosphere, overall temperature, flora
danger, many experts warn, is that	global warming will cause sea levels to rise	dramatically. Thermal expansion has already raised the
actually usable. Now take into account that	global warming can cause the sea to rise	- and as it rises, it will
the journal Science, adds to concern that	global warming may cause faster sea-level rises	than predicted, potentially increasing risks to coastal
global sea level or a rise in	ocean temperatures could cause a breakup	of the two buttressing ice shelves).
by 2100.News flash	fossil fuels cause global warming	It's kind of nitpicky, I

Fig. 4. A concordance index of "What causes what"

- **Identification of document authors**
  Although conceptually simple, in practice the identification of authors on the Web can be very difficult, if only because authentication is difficult and people can use multiple pseudonyms.

  Example: Given that online documents, including weblog posts, do not contain an explicit identification of who wrote them, the tool uses a trick to identify ownership. Most weblogs have an RSS feed, and once the RSS feed for a particular post has been identified we can rest assured that different posts with the same RSS feed are from the same author.
- **Community analysis**
  With individual authors identified, the interrelationships between people and content can be analyzed. Such analysis can become very sophisticated, using different forms and combinations of link and trend analysis.
- **Link analysis**
  Our tool provides a mechanism to identify so-called conversations. Blog conversations are sets of posts that link to each other [9]. The larger the corpus, the more useful link analysis becomes for determining the relative importance of documents (and their authors).

  Example: Jane discovers that there are few conversations in the corpus, and those that are present are not on-topic. This is unfortunate, given that online expertise often depends on the amount of links. She therefore decides to concentrate on the content of the posts.

- **Trend analysis**
  Trend analysis introduces the element of time, which can enhance interpretations significantly, especially since communities are long-term and evolving phenomena.

  Example: Given that link analysis did not uncover interesting results for the corpus, Jane decides to perform a trend analysis. First, she trends the original cue term "ice cap melting" (see Fig. 5). The trend has spikes and may point to a topic that is centered around certain events, such as the release of Al Gore's movie *An Inconvenient Truth*.

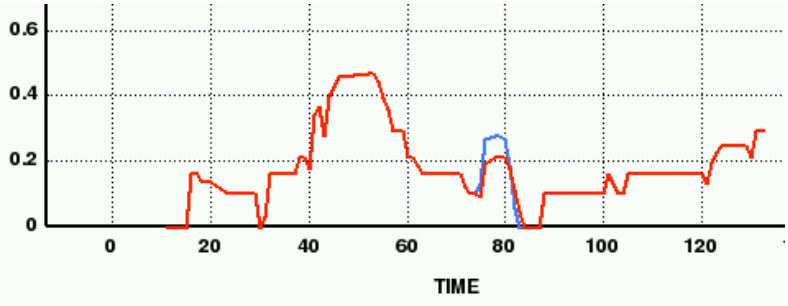

Fig. 5. Trend of ice cap melting (spikes)

Next, Jane tries a more general term "global warming" and trends it (see Fig.6). This concept seems to be on the mind of the members of the community continuously. She now considers whether "global warming" perhaps would have been a better cue term. Still, since ice cap melting is such a potent symbol of the seriousness of climate change, she decides to keep this concept for the first round of community member selection.

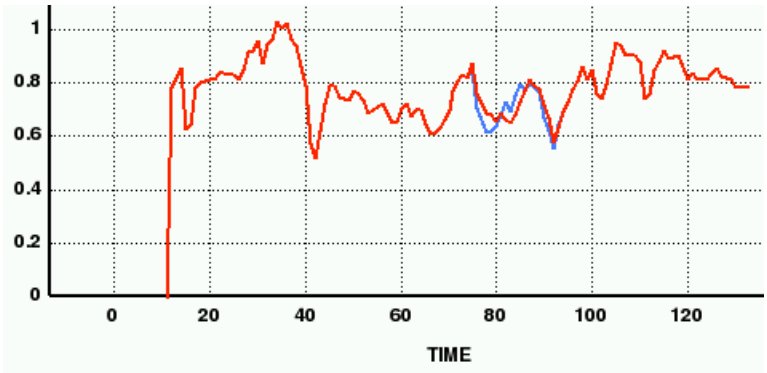

Fig. 6. Trend of global warming (continuous)

- **Ranking of document authors**
  Like for the ranking of documents, many possible ways of ranking document authors exist. tOKo provides many options to construct tailor-made rankings.

  Example: With little time left, Jane asks the tool to identify the persons who have made the most use of the concepts she has identified in her ontology (see Fig 7).

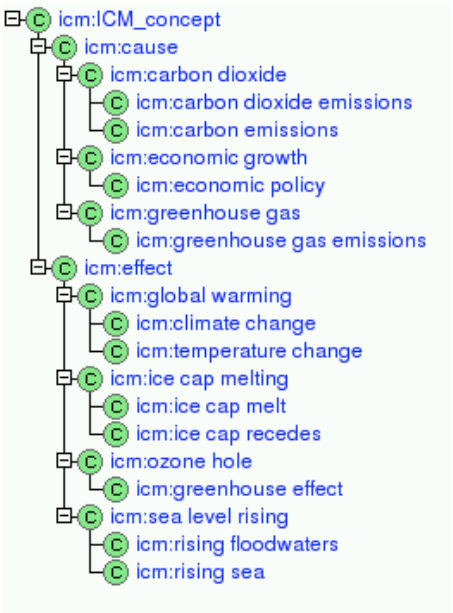

Fig. 7. Ontology of concepts relating to the causes and effects identified

- **Interpretation of ranking**
  In the sequence of steps taken, a plausibly high-quality set of authors has been identified. Given the limitations of automated analysis, a common sense human interpretation during many of these steps is inevitable. However, compared to brute force-search engine approaches, the pruned and ranked set of socio-technically determined plausible results should be of much higher quality.

  Example: The tool responds by showing a list of the bloggers who most frequently use the concepts that Jane has identified. She quickly browses the visualisations of the term clouds of the highest ranking bloggers (see Fig. 8) and then tries to contact the, in her eyes, most interesting candidates. If many bloggers rank highly, aggregate figures could be constructed, for example scatter plots with

the frequency of a selected term represented by one axis, the relevance of the term by the other axis, and bloggers represented on the plot. Such variations in visualization techniques help ensure the scalability of our approach.

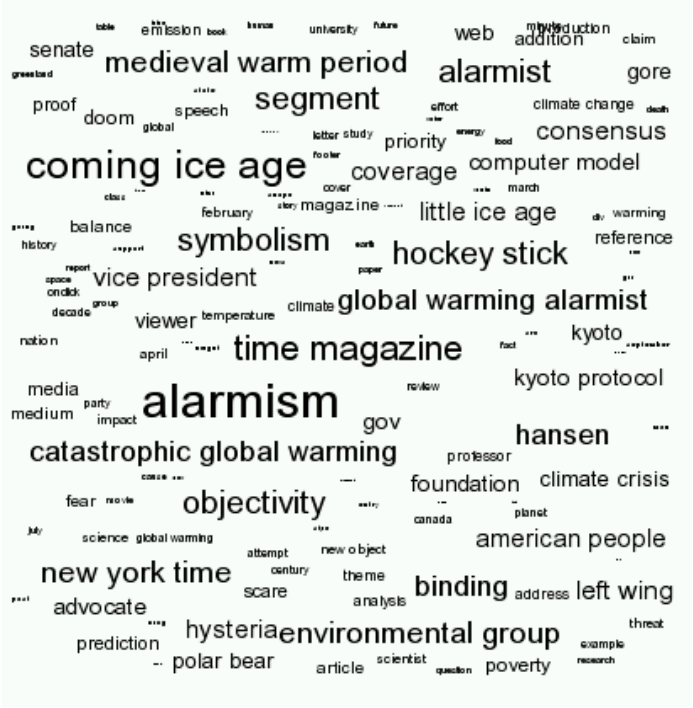

Fig. 8. Cloud of terms for a particular community member relative to other members. The bigger the term, the more this community member is using this term relative to the community.

5. Discussion

We presented only the outline of a methodology. We have shown proof of concept, which now needs to be tested in experiments with real topic communities. Of course, both the conceptualization and the implementation of the technical analysis and human interpretation steps need to be refined and expanded, in order to help this initial approach develop into a

sound and robust methodology[8]. Still, the import of the current results is that we illustrate how a *combined* approach of machine analysis and human interpretation could be a powerful way to go for community analysis. Surely, it will provide more useful results than mere computational approaches can when trying to identify expertise on fuzzy, complex topics. Along the lines of the philosophy of Doug Engelbart, whose life work has been to study the co-evolving Human System and Tool System, in particular focused on "mission-oriented communities" [10], we are interested in a process of *augmentation*, rather than automation of human intelligence. Note that the scenario involved only one person at the steering wheel, Jane. However, to make it a truly *socio*-technical approach, the communities themselves need to get involved in their own member selection. Web 2.0-like approaches could be coupled to ours, in order to make this easier to accomplish.

We used almost literal pattern searches such as `X causes ice cap melting' and `ice cap melting causes Y'. tOKo allows some very basic (concept type) ontological relations to be defined, such as 'ocean rising is-a type of sea level rising'. This allows the pattern search net to be cast wider, resulting in a larger recall. However, more complex ontological relations/networks could be introduced in the future to further increase precision and recall. For instance, the tOKo approach could be connected to the linguistically grounded, community-driven and scalable ontology engineering methodologies like the DOGMA methodology developed at STARLab [18].

In the analysis of our case, the blogosphere coverage of a document with the key causes and effects of `ice cap melting' was an important indicator to determine somebody's level of relevance for the topic community. Since the corpus contained relatively few linked posts, this additional indicator was relatively uninformative in our scenario, however. This probably has to do with this particular concept still only relatively recently having surfaced in public discourse, as indicated by the spike of posts. A term like `green house effect' both occurs in many more posts and has a much more continuous presence when analyzing its trend over time, thus being much better linked.

Scalability of analysis is an important issue. For our exploratory purpose, we downloaded only around 700 posts. To truly benefit from this analysis, many more posts need to be processed. For example, had we selected the term 'global warming', BlogPulse would have returned 70,000 items. This would have costs us many hours to download. This, however,

[8] The further development of the approach will be coordinated via the portal topixplorers.communitysense.net

can be remedied with relatively simple technical means to be provided by the blog search engines. In our case, direct access to the BlogPulse database of blog posts was offered to us for future research.

Another point is the identification of authors of documents. With blog posts this can be non-trivial. Our workaround of defining posts that come from the same RSS feed as coming from the same author is only a proxy. It shows the need for more standard approaches to community semantics, such as the ontology-based methods proposed in the Semantically-Interlinked Online Communities (SIOC)-project[9].

A related issue is that a blogger is not necessarily the creator of an idea, as many bloggers forward ideas of others. Still, they often add their own comments and insights, showing at least interest if not expertise. More importantly, many additional technical criteria and thresholds could be built into the searches. For example, if terms are mentioned in only one post, the terms are not counted. If a blogger uses many concepts and terms closely related to the focal concept, she presumably is more knowledgeable than if this is not the case. And so on.

Besides the importance of appropriate corpus, concept/term, and author selection, the impact measures themselves should become much more sophisticated. Social accounting metrics, measures of the social dimensions of online spaces, could inform the design of better indicators [17]. These could then be implemented using tOKo. The advantage of tOKo, with its five dimensions that can be combined in many different ways, together with a sequence of interpretation steps as outlined in this paper, is that very sophisticated socio-technical relevance measurement procedures and filters can be constructed that are rooted in both what can be analyzed automatically (the dimensions) and what needs human evaluation, preferably by the community as a whole (common sense).

In our analysis, we only looked at weblogs. There are many other resources which could be mined: mailing lists, news groups, web sites, databases, and so on. Still, the blogosphere is a good starting point, as blog posts are often more carefully prepared than e-mails, which are less permanent. Furthermore, in general, bloggers worldwide do perceive a shared sense of community [19], and thus could be quite willing to be invited into a topic community that recognizes their expertise. This makes it plausible that more extensive experiments along the lines we propose could produce results useful to actual problem domains like sustainable development.

Topic communities could become linking pins of potentially many different blog communities that provide them with their members. Still, many hard questions need to be answered. How can participants coming from

[9] sioc-project.org

different blog communities with different norms and cultures efficiently build an interdisciplinary topic community? [21]. A good understanding of the social norms of the contributing communities could help in better socio-technical designs of the topic community being formed [24]. Only then can (online) communities emerge that are both collaborative and adversarial, in the sense that members representing different interests can still, if not agreement, at least reach some form of consensus on what exactly they disagree on [7].

Finally, we looked only at the selection of potential community members, which is part of the initial stage of the community lifecycle. Our approach, however, could also be used to inform other stages of the lifecycle. For example, in the *Growth*-stage of a community, new people and new ideas flow into the community, and (re)combine with existing social and content structures [14]. Then, the focus could shift more to term and document analysis for its own sake, in order to deal with significant knowledge gaps.

6. Conclusion

Topic communities will become important catalysts of discussion and research on many complex interdisciplinary and societal problems. It is crucial that the most relevant experts and stakeholders can be found at the lowest possible cost. However, the generation of such communities is far from trivial, and goes beyond the capability of current political and scientific organizations. This is especially true with respect to vast and quickly evolving topics like the causes, effects and solutions of environmental and developmental crises.

In this paper, we presented a socio-technical approach for topic community member selection. It consists of a sequence of steps of human interpretation of blog analysis results provided by technical community text analysis. The result is an efficient procedure to arrive at a set of topic community candidate members likely to be knowledgeable about the topic at hand.

We only outlined this approach and illustrated its feasibility by a realistic case analysis. There is still much room for improvement, both in the conceptualization of the method steps, and the operationalization of these steps in technical analysis procedures. Still, our main goal was not to present a mature methodology, but to demonstrate that an appropriate mix of technical analysis and human interpretation steps can lead to plausibly useful results for topic community formation. Thus, instead of aiming for, in-

feasible, brute force computational approaches, our goal was to demonstrate a working combination of the advantage of computers - fast analysis of large volumes of data - with the unique strength of human beings - common sense. It is our strong conviction that such socio-technical approaches augmenting communities of human interpreters are the future of the useful mining of the wealth of community resources on the Web.

7. References

1. A. Anjewierden, R. Brussee, and L. Efimova. Shared conceptualizations in weblogs. In T. Burg, editor, *BlogTalks 2.0: The European Conference on Weblogs*, pages 110–138. Danube University of Krems, 2005.
2. A. Anjewierden and L. Efimova. Understanding weblog communities through digital traces: A framework, a tool and an example. In R. Meersman, Z. Tari, and P. Herrero, editors, *Proc. of the OTM 2006 Workshops*, LNCS 4277, pages 279–289. Springer, 2006.
3. A. Anjewierden et al. tOKo and Sigmund: text analysis support for ontology development and social research. http://www.toko-sigmund.org, 2006.
4. B.R. Barber. *Jihad vs. McWorld: How Globalism and Tribalism are Reshaping the World*. Ballantine Books, New York, 1995.
5. J. Brazelton and G. Anthony Gorry. Creating a knowledge-sharing community: If you build it, will they come? *Communications of the ACM*, 46(2):23–25, 2003.
6. J. Conklin. Wicked problems and social complexity. In J. Conklin, editor, *Dialog Mapping: Building Shared Understanding of Wicked Problems*. Wiley, Chichester, 2006.
7. A. de Moor and H. Weigand. Effective communication in virtual adversarial collaborative communities. *Journal of Community Informatics*, 2(2), 2006.
8. B. Debatin. From public/private to public privacy: A critical perspective on the infosphere. In *Proc. of DIAC-02*, May 16-19, 2002, Seattle, 2002.
9. L. Efimova and A. de Moor. Beyond personal webpublishing: An exploratory study of conversational blogging practices. In *Proc. of the 38th Hawaii International Conference on System Sciences (HICSS-38)*, Hawaii, January 2005, 2005.
10. D. Engelbart. Coordinated information services for a discipline- or mission-oriented community. In *Proc. of the 2nd Annual Computer Communications Conference*, San Jose, California, January 24, 1973, 1973.
11. P. Gongla and C.R. Rizzuto. Evolving communities of practice: IBM global services experience. *IBM Systems Journal*, 40(4):842–862, 2001.
12. R. Hirschheim and H.K. Klein. Crisis in the IS field? A critical reflection on the state of the discipline. *Journal of the Association for Information Systems*, 4(5):237–293, 2003.

13. W. Kunz and H.W.J. Rittel. Issues as elements of information systems. Technical Report 131, Institute of Urban and Regional Development, University of California, July 1970.
14. R. McDermott. Community development as a natural step: Five stages of community development. *Knowledge Management Review*, 3(5), 2000.
15. M. Roell. Business weblogs: A pragmatic approach to introducing weblogs in medium and large enterprises. In *European Conference on Weblogs*, Vienna, May 23-24, 2003.
16. S. Sassen. *Globalization and its Discontents*. The New Press, New York, 1998.
17. M. Smith. Tools for navigating large social cyberspaces. *Communications of the ACM*, 45(4):51–55, 2002.
18. P. Spyns, R. Meersman, and M. Jarrar. Data modelling versus ontology engineering. *SIGMOD Record*, 31(4):12–17, 2002.
19. N.M. Su, Y. Wang, G. Mark, T. Aiyelokun, and T. Nakano. A bosom buddy afar brings a distant land near: Are bloggers a global community? In P. van den Besselaar, G. De Michelis, J. Preece, and C. Simone, editors, *Proc. of Communities and Technologies 2005*, Milano, Italy, pages 171–190. Springer, 2005.
20. C. Verwijs, I. Mulder, R. Slagter, and H. de Poot. Positioning communities: A study on the state of the art of professional and learning communities. Technical report, Telematica Instituut, the Netherlands, 2001.
21. C. Wei. Formation of norms in a blog community. In L Gurak et al., editor, *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. 2004.
22. E. Wenger. Communities of practice: Learning as a social system. *Systems Thinker*, June, 1998.
23. E. Wenger, R. McDermott, and W.M. Snyder. *Cultivating Communities of Practice*. Harvard Business School Press, 2002.
24. B. Whitworth and A. de Moor. Legitimate by design: Towards trusted socio-technical systems. *Behaviour & Information Technology*, 22(1):31–51, 2003.