Vrije Competitie 2008
Exacte Wetenschappen

1 Project Details

1a) Project Title. Tracking News Events and their Impact

1b) Project Acronym. TNT

1c) Principal Investigator. Prof dr. Maarten de Rijke
ISLA, University of Amsterdam
E-mail: `mdr@science.uva.nl`
URL: `http://ilps.science.uva.nl/`

1d) Renewed application. This is *not* a renewed application.

2 Summary and Dutch language abstract

2a) Summary. The proposed project aims to facilitate tracking news events and determining their impact on the general public, both by professionals (media analysts, news watchers, scientists) and by the general public itself. We want to achieve this by developing algorithms that allow us to track news and measure impact largely automatically. In the scenario that we envisage, news data will be obtained from the internet, and the impact of events will be determined by analyzing the comments left behind by readers.

To be able to address our overall task, three preparatory activities are needed: (1) recognizing people, products, organizations, locations and temporal expressions, both in edited content and in user generated content, often quickly written, unedited comments left behind by readers of news messages; (2) abstracting over news stories to news incidents, clustering messages about the same incident, and summarizing the material; and (3) determining the opinions of readers on the basis of their comments.

These three subtasks have been addressed extensively in the literature, with well understood solutions. The innovative and scientifically challenging aspect of this proposal is to (1) apply them to the Dutch language; (2) apply them to noisy texts; (3) integrate them and use them to provide insights in the daily flow of news facts and the comments that they generate.

Using the solutions to the three subtasks as building blocks, we develop and test methods that will facilitate media analysis of large quantities of data. The algorithms that we aim to develop will generate well-organized, interpretable data in which the main trends and links will become visible. In sum, the project is directly aimed at finding solutions to combat the data explosion constituted by news facts and the public's responses to them. The project's results will lead to a renewed digital experience of online news.

2b) Abstract for laymen (in Dutch). Volgen van gebeurtenissen in het nieuws en het bepalen van hun invloed.

Het voorgestelde project wil het volgen van gebeurtenissen in het nieuws en het bepalen van hun invloed op het algemene publiek makkelijker maken, zowel voor professionals (media-analysten, nieuwsvolgers, wetenschappers), als het grote publiek. We proberen dit te bereiken door algoritmes te ontwikkelen waarmee we nieuws-volgen en invloed-meten in grote mate kunnen automatiseren. Nieuws betrekken we van het internet en de invloed van gebeurtenissen in het nieuws bepalen we door de reacties die lezers achterlaten bij nieuwsberichten te analyseren.

Voor deze taak zijn drie voorbereidende activiteiten nodig: (1) het herkennen van personen, producten, organisaties, locaties en tijden zowel in "nette" teksten geschreven door professionals als in zogenoemde *user generated content* (UGC), snel geschreven, niet nagekeken teksten afkomstig van het lezerspubliek; (2) het abstraheren van nieuwsverhalen naar gebeurtenissen, berichten over dezelfde gebeurtenis kunnen clusteren en die gebeurtenissen automatisch kunnen samenvatten; (3) het bepalen van meningen van lezers op basis van hun reacties.

Afzonderlijk zijn deze drie taken al uitvoerig onderzocht en er bestaan goed geëvalueerde oplossingen. Het innovatieve en wetenschappelijk uitdagende aan dit project is om (1) ze toe te passen op het Nederlands, (2) ze toe te passen op ruizige teksten (UGC), en (3) de drie taken te integreren en ze te gebruiken om inzicht te bieden in de dagelijkse stroom van nieuwsfeiten en de enorme hoeveelheden reacties die daarop volgen.

Met de oplossingen voor deze drie taken als bouwstenen ontwikkelen en testen we methoden waarmee media-analyse van enorme hoeveelheden data mogelijk (want geautomatiseerd) wordt. De te ontwikkelen algoritmen genereren geordende, interpreteerbare gegevens waarin de grote lijnen en verbanden inzichtelijk worden. Het onderzoek in dit project richt zich dus rechtstreeks op het vinden van oplossingen om het hoofd te bieden aan de data explosie van nieuwsfeiten en alle reacties daarop. De resultaten van het onderzoek zullen leiden tot een hernieuwde digitale beleving van online nieuws.

3 Classification

The project falls within the discipline of *Computer Science*. Relevant *Nationale Onderzoeksagenda Informatie- en Communicatietechnologie 2005-2010 (NOAG-ict)* research themes: **3.2 Data Explosie** (ICT-disciplines: Algorithms and Computation Theory; Hypermedia, Hypertext and Web; Information Retrieval; Knowledge Discovery in Data) and **3.3 Digitale Beleving** (ICT-disciplines: Computer-Human Interaction; Information Retrieval).

4 Composition of the Research Team

Name	Title	Role	Expertise	Affiliation
de Rijke, M.	Prof.dr.	Applicant	Information storage and retrieval	ISLA, U. Amsterdam
Marx, M.	Dr.	Co-applicant	Semi-structured data	ISLA, U. Amsterdam
Ahn, D.	Dr.	Advisor	Event modeling	Powerset, S.F., USA
Franz, R.	Drs.	Advisor	Media analysis	TrendLight, Amsterdam
Moens, F.	Dr.	Advisor	Information extraction and summarization	Kath. U. Leuven
Steinberger, R.	Dr.	Advisor	Language technology/news search	European Commission Joint Research Centre (JRC)
NN		Ph.D. student		ISLA (to be funded by NWO)

Maarten de Rijke is professor of “Information processing and internet” at ISLA. He works on information retrieval and language technology aspects of access to online content, and has a strong interest in user generated content. Maarten Marx is assistant professor in computer science at ISLA, specialized in semi-structured information and XML. He works on foundational issues concerning expressivity and complexity of XML query languages, query-rewrite systems, access control and data mediation. He has a more practical research interest in focused information retrieval from XML documents, exercised both within the INEX initiative and in building search engines for political data for the Dutch and Belgium parliamentary elections. The Ph.D. student (AIO) will be located at ISLA and supervised by De Rijke and Marx.

Advisors have been included based on the workplan and the required expertise. Because an important aspect of the proposal is to work on Dutch data the team contains 4 native Dutch speakers and Dr Ahn who speaks Dutch fluently. From 2005 till mid 2007 David Ahn was a postdoc at ISLA working on temporal information extraction; he is presently a natural language scientist at Powerset. He will provide input and advice on clustering and representing events (WP4; see below). Raymond Franz is director of TrendLight Netherlands B.V., a company specialized in media analysis, working for companies, governmental and non-profit organizations. TrendLight is a problem-holder, a professional organization that needs computational support in tracking events and their impact; TrendLight will provide worked out use-cases and hand-labelled training data, and give advice at “design choice moments” (schema design in WP2, sentiments in WP5). Francine Moens is professor in computer science at the Katholieke Universiteit Leuven in Belgium, doing research in text based information retrieval, information-extraction, summarization and text mining. She worked on coreference resolution in Dutch and will provide input and advice on WP3. Ralf Steinberger is one of the driving forces behind EMM NewsExplorer, which generates daily news summaries and allows users to compare reports on the same event across languages. He will provide input and advice on clustering and representing events (WP4).

5 Research School

The research will be embedded in the Dutch Research School for Information and Knowledge Systems (SIKS).

6 Description of the Proposed Research and Application Perspective

6a—Description of the Proposed Research

a—Research Questions and Desired Results

News is of interest to both the general public and a broad range of professionals, including economists, marketeers, information and media-analysts. With the web, the amount of news has been increasing exponentially. It is not feasible for users to go through the information without some form of pre-processing and organizing. But there's more. Online news-stories are not isolated text snippets—they are richly linked to the online world around them, with intrinsic connections to other stories and (implicit) links to background information, to the entities featuring in them. In recent years, online news has acquired yet another dimension: increasingly, news-sites allow their readers to leave behind *comments* that are at least as interesting and valuable as the triggering news-events themselves since they provide direct and near real-time information about the impact of events.

Online news-stories, richly linked in the manner sketched above are the objects being studied in **TNT**. We aim to develop, implement and test methods for tracking and understanding them. The approach of **TNT** is to automatically create links between stories and organize the linked information. We foresee three groups of linked information:

- *Stories*: There is an overwhelming repetition of news-stories covering the same event. We cluster these into events [8]. Collections of events are modelled as hierarchically and chronologically linked threads of events [17, 38].
- *Background information*: Almost every news-story answers the four main questions in journalism: *who*, *when*, *where* and *what*. Each story contains implicit links to named entities (NEs) representing actors and locations, to timepoints and periods, and to topics. Making these links explicit and incorporating them in rich event models has been shown to improve accuracy in tasks as *new event detection* [27, 28, 52].
- *Comments*: Comments are *user generated content* (UGC). Linked to news-stories, they provide a connection between news-events and the public's opinion [5], and thus allow us to efficiently and reliably quantify the impact of the event, across multiple sources, and to determine the perspectives that emerge.

To provide focus and clear annotation and evaluation criteria, the project is aimed at one specific type of user: the professional media analyst, who provides an analysis of the formation, spreading and development of news, and describes its impact on specified target audiences. Here's an example use case:

Example use case Postbus 51 (Dutch governmental public information agency) plans a campaign against violence at schools triggered by recent fatal incidents in schools. They want to use comments on news-stories covering these incidents to obtain a better picture of their target population. This is only possible if comments from many different sites are harvested, stored, brought into a uniform format, grouped by event, enriched with links to related stories and entities, and made accessible for aggregation and analysis.

We propose a research program with *three main goals*: 1. to collect the bulk of this data (news-story, comment-thread pairs); 2. to develop algorithms that help make the data accessible to professionals (by clustering, aggregation, summarization, linking to background information, and linking news into event-threads) ; and 3. to develop algorithms that enable professional media analysts to track news and determine its impact. We limit the project to collecting Dutch data, but in order to compare and disseminate our results, techniques will be developed mostly language-independently. To achieve these goals we need to make progress beyond the state-of-the-art in three areas: *semistructured data* (SSD), *information retrieval* (IR), and *language technology* (LT). Specifically, the project addresses six *major research challenges*:

- *IR-RQ1*: Tracking news events and their impact as measured by comments to news-stories about these events.
- *IR-RQ2*: Efficient and robust recognition and reconciliation of NEs in news-stories and related comments.
- *LT-RQ3*: Clustering news-stories into events and events into event-threads.
- *LT-RQ4*: Robust sentiment analysis in comments on news-stories.

- *SSD-RQ5*: Schema mapping and data integration in a dynamic setting.
- *SSD-RQ6*: Transparent archiving in XML.

We have three kinds of *desired results*:

- *Data collection*: a fairly complete, normalized, well-described, automatically annotated and enriched archive of Dutch pairs of news-story summary and comments; together with a robust data-collection infrastructure for maintaining the archive.[1]
- *Algorithms and tools*: a set of efficient and scalable tools for text analysis and aggregation of news-story and comments pairs, with an emphasis on Dutch. The tools link stories to background information, interconnect related stories into event-threads, and link events to sentiments.
- *A prototype workbench*: a workbench for information and media analysis on news-events and their impact. The purpose of the workbench is to obtain feedback from the intended users and to learn from the application of our tools to concrete problems.

Viewing online (textual) news-stories as anchored in a network of stories, entities and comments, **TNT** combines theoretical, experimental and applied research to create aggregation and analysis methods that support efficient and effective tracking of news-stories and their impact.

b—Approach and Methodology

Before presenting our approach, we highlight two important issues that affect our work. Following [44], we view theoretical, experimental, and applied aspects of academic research as inextricably interlinked, where each builds on, and feeds into the other. To facilitate the desired interaction, the project aims to have a baseline prototype at an early stage; this is possible because the host institute has baseline versions of key-components; see §8. Second, UGC and edited content differ significantly. First, language use in UGC diverges from that in edited content. Second, unlike mainstream media, UGC often refers to private experiences and this will contribute to higher referential ambiguity of names and other entity mentions, thus complicating the linking that we foresee. Third, our comments have structure that news articles do not—this may provide opportunities [35].

Our approach to the *SSD* aspects of **TNT** is to ground it in the theory and practice of data integration and mediation. We use the schema mapping and data integration techniques incorporated in the IBM-Clio- and the IBM-Information Server systems [20, 34] to integrate the many news sources that we cover into one target schema. Because news-sites change, come and go, we extend these to a dynamic setting, with emphasis on maintaining an efficient and understandable (integrated) target schema.

The normalization and annotation efforts in **TNT** result in large amounts of document-centric and data-centric data, stored in several formats (relational DB, text files, RSS, (X)HTML, special purpose XML). We use data mediation to develop a transparent mapping from our data to an XML Schema [6, 9]. The novel aspect here is the heterogeneity of the sources. The lack of query languages for combinations of heterogeneous sources [16] is a challenge; Datalog with predicates corresponding to materialized views seems a viable option [19].

Our approach to the *LT* aspects of **TNT** is based on semi-supervised learning. For RQ3, recent work on *incident threading* [17] forms our starting point. **TNT** will adopt a two-stage approach to forming incident networks [17], building on a hierarchical agglomerative clustering algorithm with sampling [47], and port this to our Dutch language setting—this approach scales better than other methods considered and provides accurate clusters. We expect that clustering will be improved by using comment-threads as these provide the redundancy required by statistical methods. For development and testing an annotated corpus will be created [17].

For RQ4, we aim to identify sentiment at the comment level. We focus on two types of sentiment: *criticism* and *support* [50]. We will use the sentiment analysis tools being developed for news within the NWO-STEVIN funded DuOMAn project [14]. The differences in language use between UGC and news mean that models trained on news will be less effective on comments (as noted at the TREC blog track [41]). Second, the expected high degree of referential ambiguity will complicate aggregation. Third, the volume of content makes result presentation especially important. Media analysts still need to delineate the range of information they would like to extract from news-comments, but we do know the basic tasks. The

[1] For copyright reasons, we only archive summaries plus permalinks to sources; comments are archived in full.

first will be to re-train the DuOMAn sentiment classifier. The resulting sentiment mining task is similar to the TREC opinion retrieval task, and we will experiment with targeted opinion mining, combining retrieval and opinion classification [41].

The *IR* challenges of the project are associated with RQ1 and RQ2. Concerning RQ2, key ingredients of an event are the answers to *who, where, when* questions. We collect these using a home-grown Dutch NE recognizer, which works well on edited texts [46]; it uses machine learned classifiers to identify names and determine entity type; we will adapt the recognizer to comments. Recognized NEs need to be reconciled, a task consisting of disambiguation and normalization [13]. [12] describes a method for addressing ambiguity and synonymy, and applies it to news-stories; this is improved upon in [4] by using Wikipedia entries as canonical forms [33] and coreference theory [51] for mapping variants to canonical forms. We build upon [4] and supplement it with disambiguation-methods based on language models and string-similarity methods, combined using standard voting-based techniques.

Our research efforts come together in our approach to RQ1. We take *impact* to refer to two aspects of the comments: quantitative (how many, how frequent) as well as qualitative (aspects covered, dominant sentiments). Based on prior experience in predicting and explaining mood levels and bursts in blogs [11, 37], and observing temporal regularities in news-comments data [5] we believe that quantitative regularities in the comments may be uncovered using regression-type analysis. We capture qualitative aspects of news-comments based on summarizing and aggregating news (bringing in linked entities and stories), and then summarizing the comments based in part on corpus comparisons (to the news, so as to reflect credible language usage) and on clustering similar sentiments [49, 50], following the method of [18]. Within each cluster, we rank comments based on relevancy to the triggering story, opinionatedness, and credibility [42, 48].

c—Significance and Urgency

A rapidly growing portion of the web consists of UGC. UGC contains valuable information that is difficult to extract with tools developed for and trained on edited texts [35]. Initial experiments and a pilot investigation show that data extraction from UGC in the form of comments to news-articles is feasible. [5].

The project is *significant* for two NOAG-ict research themes: *data explosion* and *digital experience*: news is, and increasingly will be, consumed online, with the attached comments forming an integral part of the news-story. The amount of online news is already overwhelming, and a single news-story can have an overwhelming number of comments. This is a new digital experience which needs intelligent software.

The project is *urgent* for two reasons: 1. Comments contain valuable information for media and information analysts, but the sheer volume prevents large-scale quantitative approaches. The tools developed in **TNT** facilitate this. 2. Whereas (commercial) news-articles are properly archived in the Netherlands, and available through LexisNexis, the situation with comments to these articles is unclear. The tools developed in **TNT** (esp. in WP2 and WP7 below) will show that properly archiving news-comments is feasible.

d—Related Work

The proposal combines techniques from three core areas in web-information technology: *SSD*, *IR*, and *language technology* (LT). We group the related work by these areas. Within SSD the most important related work concerns data integration, data exchange [20, 29] and data mediation [30]. For data integration, standardization of data-values is a basic but crucial topic, and this links it with LT and IR [24]. A key challenge is entity resolution (aka data deduplication): determine if two data objects refer to the same real-world entity [13, 26]. **TNT** extends this semantic resolution of entities to events, a new topic in data integration. This relates to the event clustering algorithms developed within TDT [8].

News-events have been of interest to researchers in IR and LT for a long time [31]. There are multiple perspectives on events that crop up in the literature: linguistic semantic approaches that view events as any changes of state that are somehow conceptualized (lexically, grammatically, etc.)—TimeML's events fall into this category, and much work on semantic role labeling is related [45]. Second, there is the Topic Detection and Tracking view, where events are an explicit motivation but are not really modeled [8]. And, third, there is the MUC/ACE approach, which is somewhere in between [7]. Our core research question (RQ1) is related to the TDT work mentioned earlier, and to information fusion as described in [10] and "practised" at WiQA [21] and WebCLEF [22].

Sentiment analysis is an LT task that is increasingly popular; see the earlier references in §6.a.b. Aggregating and tracking sentiments over time is a task being addressed at the NTCIR-MUST workshop on multimodal summarization for trend information [40], while it will also be addressed at the new question answering track at the Text Analysis Conference (TAC) to be hosted by NIST (November 2008; De Rijke is a member of the advisory committee for TAC).

e—Embedding

TNT continues research on dynamic, news related UGC currently conducted by the host institute, of which we have so far covered many aspects, including IR [15, 35], evaluation [41], entity normalization [4], event extraction [1], semi-structured information [2, 3, 5, 25], credibility [48], sentiment analysis [37], cross-channel tracking of news and its perception [11, 23], and user needs concerning UGC [36].

These approaches achieved competitive results and are brought together in **TNT**. In addition, our team has expertise with running large-scale demonstrators (MoodViews.com, VerkiezingsKijker.nl).

The host institute is one of the largest academic research groups in IR in Europe; it participates in, and helps organize, worldwide evaluation efforts such as TREC, CLEF, INEX, and NTCIR, and co-organized SIGIR-2007.

6b—Application Perspective

The project has potential for applications in several domains:

Academic Our research group needs application-driven projects; the present proposal is part of its theory-experiment-application work cycle.

Databases Data integration tools start becoming part of commercial DBMS's (e.g., within IBM's DB2). Our work on entity and event recognition and normalization is applicable in the standardization toolkit of a DBMS.

Intelligence Intelligence organizations are interested in tracking responses to news-events, especially if this can be complemented with demographic analysis (to determine who the commenters are).

Media-analysis TrendLight is currently using news-corpora based on written editions. Extending these with web editions and with the additional comments on news-stories by large numbers of web users will make measuring *the effect* of media on the public easier, faster, cheaper and more scalable.

News tracking News search engines already cluster news stories based on textual content. Lifting this to tracking news events and combining stories into news timelines will yield a new, and we believe, desirable user experience. Content analysis on newspapers which is currently done manually (e.g., for Dutch by *Nieuwsmonitor* [43]) could be automated and scaled to continuous large-scale coverage.

TNT has applications as part of new search and discovery technology being demanded by the dynamic nature of the web. *Tracking* is a natural information need in a dynamic environment; *summarization* and *impact determination* help cope with information overload.

7 Project Planning

7a—Project Structure

The project will run for 4 years starting in the fall of 2008. The project is structured into 8 work packages (WPs): a preparatory data-collection WP, 6 WPs corresponding to the research challenges RQ1–RQ6, and the final thesis writing. Figure 1 shows the project's organization and Figure 2 provides an overview of its WPs and their planning. We invest heavily in the data-collection, -preparation and -integration phases (RQ2–RQ5), creating a solid foundation for the project's main task: "tracking news and its impact" (RQ1). The data-archiving task (RQ6) closes the

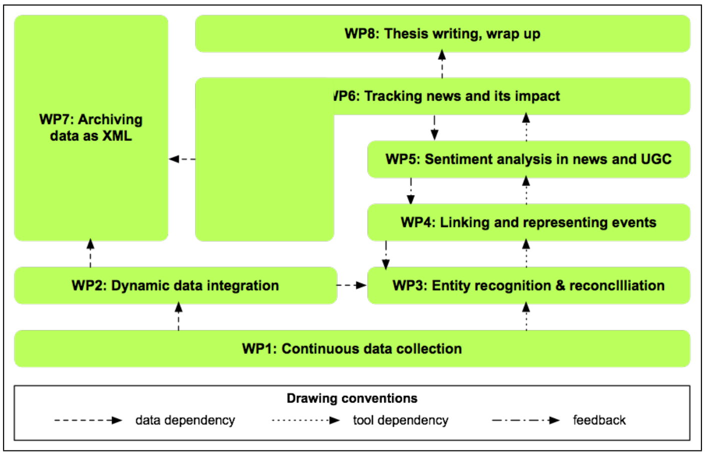
WP8: Thesis writing, wrap up
WP7: Archiving data as XML
/P6: Tracking news and its impact
WP5: Sentiment analysis in news and UGC
WP4: Linking and representing events
WP2: Dynamic data integration
WP3: Entity recognition & reconclliation
WP1: Continuous data collection
Drawing conventions
data dependency
tool dependency
feedback

Figure 1: Main stages of the project.

research tasks. Every WP connected to a research challenge will result in a deliverable submitted to a suitable workshop, conference, or journal.

The project will be developed in four R&D phases, each leading to a milestone and targeting the completion of (the implementation of) one or more specific WPs; see Figure 2. In the start-up phase the emphasis will be on platform and basic infrastructure development. **Milestone I** (month 9) is reached when core data collection and integration facilities are available. **Milestone II** (month 24) will incorporate modules developed in WP3–WP5; the focus of the milestone will be on core ingredients for news and impact tracking. During the next phase, leading to **Milestone III** (at month 36), we integrate the components so as to realize a tracking prototype.[2] In each R&D phase, the quality of the tracking platform will be improved and extended, and where needed made more robust or efficient.

WP	**RQ**	**WP title**	**Main development period**	**Start/end month**	**Milestone(s)**
WP1		Continuous data collection	1–6	1/48	I
WP2	RQ5	Dynamic data integration	3–9	3/48	I
WP3	RQ2	Entity recognition & reconciliation	9–12	9/48	II
WP4	RQ3	Linking and representing events	13–18	13/48	II
WP5	RQ4	Sentiment analysis in news and UGC	18–24	18/48	II
WP6	RQ1	Tracking news and its impact	24–36	24/48	III
WP7	RQ6	Archiving data as XML	36–40	36/40	IV
WP8		Thesis writing, wrap up	40–48	40/48	IV

Figure 2: Work package overview and planning

7b—Work Packages and their Risks

WP1. Collect Dutch language (news-story, comment-list) pairs from newspaper sites. Extend to sites that offer the same (news-item, comment-list) structure. Obtain extensible and robust data collection infrastructure that will run for duration of project. *Risk*: Sites objecting to being crawled and analyzed. *Remedy*: Obey robot standards and customs. If that does not help, remove them from our list. In worst case, fall back to data already obtained [5].

WP2. Integrate (news-story, comment-list)-pairs data from multiple sources and store in uniform format: a solved problem for news-stories (by RSS-technology); not so for comments. For impact measurements, collect metadata from comments (e.g., commenters' location). *Risk:* Too much variation in schemas to handle automatically. *Remedy:* Stick to a core of similar sources, which can be integrated by hand, if necessary.

WP3. Create accurate links from news stories and comments to entities. Need accurate (recall-oriented) mappings of entities mentioned in comments to entities in triggering news-story. *Risk:* Too creative language use. Unknown actors. *Remedy:* Focus on events having well-known actors (e.g., Dutch politics, soccer, show business).

WP4. Identify relations between news-stories: same incident, precondition, consequence, etc. [17]; investigate potential of comments to improve relation detection. *Risk*: Not enough training data. *Remedy:* Semi-supervised learning methods, supplemented with rule-based training data creation [32].

WP5. Port sentiment analysis tools to be developed within the DuOMAn project to user generated comments, identify targets and attitudes towards those targets. *Risk:* Too much noise. *Remedy:* Fall back to "credible" comments that feature high quality language usage [48]. *Risk*: DuOMAn does not deliver sentiment analysis tool kit. *Remedy*: Port in-house opinion mining tools developed for TREC blog track.

WP6. Create quantitative impact measurement tools based on regression. Create qualitative measurement tools based on clustering and in-house information fusion tools. *Risk:* Data too noisy for regression tools to perform well. *Remedy:* Perform data cleaning first. *Risk:* Data too noisy for information fusion tools to perform well. *Remedy:* Perform data cleaning first.

WP7. Store our work for later use. Extend XML data-mediation to the case with multiple heterogeneous sources. *Risk:* Formats and sources too heterogeneous. XML is not adequate for the complexity of the data. *Remedy:* Restrict the sources mapped to XML. *Risk:* Not allowed to store copyrighted (news) data. *Remedy:* Restrict to storing and mediating summarized data.

[2] A requirement analysis concerning media analysts' usage of UGC in addition to their current workflow is not part of **TNT**; this analysis is currently conducted by ISLA and TrendLight.

7c—Training and Education

Training and education are aimed at developing scientific expertise, and acquiring professional competencies. The PhD. student should become able to fully understand, critically analyze, and contribute to research at the frontiers of science. Supervision and training at the host institute are governed by several instruments. The student has two supervisors; supervisors and student draw up a training and education plan, which is evaluated annually. Students follow graduate courses locally, within the national research school, and at international summer schools. A working visit to a foreign university, research institute, and/or industrial research group is a key ingredient of the plan.

8 Expected Use of Instrumentation

8a—In-house Resources

We list baseline versions of software and evaluation resources available in-house, together with the WP in which they will be used: web crawling, scraping and cleaning infrastructure: RSSManager and SSScrape (used in WP1); database schema for storing and representing news, comments and metadata [5, 39] (used in WP2, 7); NE recognizer for Dutch [46] and NE normalization algorithm for English [4] (used in WP3); sentiment analysis tools to be developed with the DuOMAn project [14] (used in WP5); multiple TREC, INEX and CLEF evaluation scripts and test sets, Wikipedia dumps in XML and relational database format, and lexicons based on Dutch news-corpora (used in WP3, 4, 5, 6); finally, standard machine learning and language modeling toolkits will be used (Weka, SVM-Light, Lemur, Indri, etc).

8b—Additional Storage or Processing Equipment

We do not request any additional storage or processing equipment.

9 Literature

Five most relevant publications by applicants

[1] D. Ahn, J. van Rantwijk, and M. de Rijke. A cascaded machine learning approach to interpreting temporal expressions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 420–427, April 2007.

[2] I. Fundulaki and M. Marx. Mediation of XML Data through Entity Relationship Models. In *Proceedings of the First International Workshop on Semantic Web and Databases*, 2003. In Conjunction with VLDB.

[3] V. Jijkoun and M. de Rijke. Retrieving answers from frequently asked questions pages on the web. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 76–83, New York, NY, USA, 2005. ACM Press.

[4] M. Khalid, V. Jijkoun, and M. de Rijke. The impact of named entity normalization on information retrieval for question answering. In *30th European Conference on Information Retrieval (ECIR 2008)*. Springer, 2008.

[5] A. Schuth, M. Marx, and M. de Rijke. Extracting the discussion structure in comments on news-articles. In *9th ACM International Workshop on Web Information and Data Management (WIDM 2007)*, pages 97–104, 2007.

Other references

[6] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the web*. Morgan Kaufman, 2000.

[7] ACE. Automatic content extraction, 2008. `http://www.nist.gov/speech/tests/ace/`.

[8] J. Allan, editor. *Topic Detection and Tracking:Event based Information Organization*. Kluwer Academic Publishers, 2000.

[9] B. Amann, C. Beeri, I. Fundulaki, and M. Scholl. Querying XML sources using an ontology-based mediator. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 429–448. Springer-Verlag, 2002.

[10] E. Amigo, J. Gonzalo, V. Peinado, A. Penas, and F. Verdejo. An empirical study of information synthesis tasks. In *Proceedings ACL 2004*, 2004.

[11] K. Balog, G. Mishne, and M. De Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *Proceedings 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, April 2006.

[12] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *EMNLP-CoNLL '07*, pages 708–716, 2007.

[13] X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 85–96, New York, NY, USA, 2005. ACM.

[14] DuOMAn. Dutch Language Online Media Analysis, 2008. `http://www.nwo.nl/projecten.nsf/pages/2300141067`.

[15] B. Ernsting, W. Weerkamp, and M. de Rijke. The University of Amsterdam at the TREC 2007 Blog Track. In *TREC 2007 Working Notes*, November 2007.

[16] W. Fan, F. Geerts, and F. Neven. Expressiveness and complexity of XML publishing transducers. In *PODS '07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 83–92, New York, NY, USA, 2007. ACM.

[17] A. Feng and J. Allan. Finding and linking incidents in news. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 821–830, New York, NY, USA, 2007. Acm.

[18] S. Fissaha Adafre, V. Jijkoun, and M. de Rijke. Fact discovery in Wikipedia. In *2007 IEEE/WIC/ACM International Conference on Web Intelligence*, November 2007.

[19] G. Gottlob, C. Koch, R. Baumgartner, M. Herzog, and S. Flesca. The Lixto data extraction project: back and forth between theory and practice. In *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–12, New York, NY, USA, 2004. ACM.

[20] L. M. Haas. Beauty and the beast: The theory and practice of information integration. In T. Schwentick and D. Suciu, editors, *ICDT*, volume 4353 of *Lecture Notes in Computer Science*, pages 28–43. Springer, 2007.

[21] V. Jijkoun and M. de Rijke. Overview of the WiQA task at CLEF 2006. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 265–274, 2007.

[22] V. Jijkoun and M. de Rijke. Overview of WebCLEF 2007. In A. Nardi and C. Peters, editors, *Working Notes for the CLEF 2007 Workshop*, September 2007.

[23] V. Jijkoun, M. Marx, M. de Rijke, and F. van Waveren. Electoral search using the VerkiezingsKijker: an experience report. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1155–1156, New York, NY, USA, 2007. ACM Press.

[24] T. Johnson and T. Dasu. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003.

[25] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Structured queries in XML retrieval. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 4–11, New York, NY, USA, 2005. ACM Press.

[26] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 802–803, New York, NY, USA, 2006. ACM. ISBN 1-59593-434-0. doi: http://doi.acm.org/10.1145/1142473.1142599.

[27] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304, New York, NY, USA, 2004. ACM Press.

[28] G. Kumaran and J. Allan. Using names and topics for new event detection. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 121–128, Morristown, NJ, USA, 2005. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1220575.1220591.

[29] M. Lenzerini. Data integration: A theoretical perspective. In *PODS*, pages 233–246, 2002.

[30] A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *VLDB '96: Proceedings of the 22th International Conference on Very Large Data Bases*, pages 251–262, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1-55860-382-4.

[31] I. Mani and G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, pages 69–76, 2000.

[32] B. Medlock and T. Briscoe. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 992–999, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[33] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA, 2007. ACM.

[34] R. J. Miller, M. A. Hernández, L. M. Haas, L. Yan, C. T. H. Ho, R. Fagin, and L. Popa. The Clio project: managing heterogeneity. *SIGMOD Rec.*, 30(1):78–83, 2001.

[35] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, 2007.

[36] G. Mishne and M. de Rijke. A study of blog search. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *Advances in Information Retrieval: Proceedings 28th European Conference on IR Research (ECIR 2006)*, volume 3936 of *LNCS*, pages 289–301. Springer, April 2006.

[37] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. In N. Nicolov, F. Salvetti, M. Liberman, and J. Martin, editors, *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, pages 145–152. AAAI Press, March 2006.

[38] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 446–453, New York, NY, USA, 2004. ACM Press.

[39] NewsComments. `http://zookma.science.uva.nl/newscomments`, 2008.

[40] NTCIR. Evaluation of information access technologies, 2008. `http://research.nii.ac.jp/ntcir/`.

[41] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*. NIST, 2007.

[42] V. Rubin and E. Liddy. Assessing credibility of weblogs. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (CAAW)*, 2006.

[43] O. Scholten and N. Ruigrok. Politiek en politici in het nieuws in vijf landelijke dagbladen. Stichting Het Persinstituut, De Nederlandse Nieuwsmonitor, `http://www.nieuwsmonitor.net/continu/politiek_2005.pdf`, 2006.

[44] D. E. Stokes. *Pasteur's Quadrant: Basic Science and Technological Innovation*. Brookings Institution, 1996.

[45] TimeML. Markup language for temporal and event expressions, 2008. `http://www.timeml.org/`.

[46] E. F. Tjong Kim Sang. Memory-based named entity recognition. In *Proceedings of CoNLL-2002*, pages 203–206. Taipei, Taiwan, 2002.

[47] D. Trieschnigg and W. Kraaij. Scalable hierarchical topic detection: exploring a sample based approach. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 655–656, New York, NY, USA, 2005. Acm.

[48] W. Weerkamp and M. de Rijke. Credibility improves topical blog post retrieval, 2008. Submitted to *ACL 2008*.

[49] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Comput. Ling.*, 30(3), 2004.

[50] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3), 2005.

[51] F. Wu and D. S. Weld. Autonomously semantifying wikipedia. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *CIKM*, pages 41–50. ACM, 2007.

[52] K. Zhang, J. Zi, and L. G. Wu. New event detection based on indexing-tree and named entity. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 215–222, New York, NY, USA, 2007. ACM Press.

10 Requested Budget

`<Omitted>`