COMMIT

PROJECTPLAN

WORKPACKAGES

DELIVERABLES

BUDGET

RETRIEVAL FOR INFORMATION SERVICES (P01)

Projectleader prof.dr. Maarten de Rijke, Universiteit van Amsterdam

1. Background

To achieve their goals, organizations, individuals and devices alike need to be aware of key players, events and attitudes in the world around them that may affect them or their interests. This is true for financial organizations surviving in an unstable economic environment, for governmental organizations with responsibilities to monitor open sources, for knowledge-intensive companies tapping into the collective wisdom of their employees, and for an individual aiming to engage in cultural activities.

The project shares with MultimediaN its goal of providing semantic access to large repositories of textual and multimedia information in general and to semi-structured information in particular. To this it adds robust language technology, analysis of subjective information, analysis of dynamic information, large-scale information extraction, distributed data processing for online text analysis, and interactive reporting and visualization facilities.

The COMMIT themes are: I-services (news, culture, and databanks), Science (generic, beta), User aspects (visualization, sensing, and interaction) and Analysis (methods, mining, and fusion).

The project will develop, and enable the exploitation of, open source and open standards tools to support *semantic search:* retrieval models, algorithms and implementations that combine information retrieval and information extraction technology to provide genuine information access that goes beyond standard document retrieval—ranking and characterizing entities (such as people, organizations, products, …) based on their profiles, on the relations and developments they are part of, and the attitudes in which they are engaged. The project's thematic lines start from existing toolkits and go through multiple cycles of use case analysis, algorithm development, methodological grounding and tool professionalization steps. A range of content and problem owners (ANP, Internet Recherche Netwerk, Ilse Media, NIBG, Persgroep, TrendLight) provide focus and grounding by supplying the project with use cases. The project's solution partners (EUVision, Internet Recherche Netwerk, Talking Trends, Teezir, Textkernel, WCC) will initiate exploitation during its lifetime.

A particular challenge for semantic search is the robustness and coverage of text and multi-media information extraction techniques [21]. Techniques from open domain information extraction like [20] need to be integrated and specialized to economic and political contexts. Interactive tagging has been explored [23], but is still in its infancy. Current approaches to issue management are neither automated nor semantic; existing toolkits are limited to simple counts and relevant linguistic resources are absent. The required algorithms for sentiment analysis and for determining associations between issues and stakeholders are limited to static information [1]: today's text understanding tools are ill-equipped for highly dynamic domains, having been developed without adaptation in mind [2], with a focus on edited content (as opposed to user generated content [3]), and largely ignoring social and subjective aspects of content [4]. Most

text summarization methods to date are limited to extractive summaries [5], but there is growing interest in automatic information presentation [6, 7]. Framing has so far not been given a computational treatment, neither on text nor on multimedia content [8]. Cross-media mining of textual and numeric data rarely involves semantic features on the textual side; see, e.g., [17]. Speech recognition technology has matured sufficiently to provide an adequate basis for robust spoken content search systems, but such systems have not been semantified yet [22]. Although contextualized content-based video analysis has left its infancy, current solutions are limited and unable to deal with heterogeneous pictorial data or with highly subjective aspects [9, 10]. The challenge of recognizing sentiment in video material from body signals has only recently been identified and is far from a solved problem [16, 24].

2. Problem description

In our networked society and information infrastructure more relevant information, answers to hard questions, experiences, opinions and media are potentially available in just a few mouse-clicks than ever before. Many organizations are faced with the task of having to identify key players and monitor developments in a complex information landscape, with a clear need for structuring and aggregation of information into actionable insights. For example, the amount of shareholder value attributable to companies' brand reputations is enormous, often in the billions. The value of the brand reputations can represent between 20–70% of a company's market cap. To help protect brand reputations, organizations need to anticipate, recognize in near real-time and take appropriate action on emerging trends and issues that are likely to affect them [11]: for financial and societal organizations alike effective management of political and economic issues is especially urgent. A second example is provided by knowledge intensive organizations that aim to tap into the collective intelligence of their workforces. Here, employees' interactions in discussion forums and other social media need to be observed with the aim of developing new strategies, services and products. A third example is provided by emerging intelligent devices: the issue management task faced here is to be semantically aware (of location, of social relations, of emotions, etc), adaptive and responsive to subjective aspects of the people using them. Due to a lack of appropriate tools it is difficult and cumbersome for financial institutions as well as high-tech and other knowledge-intensive organizations to engage in effective real-time issue management.

The project addresses the challenge of identifying and aggregating diverse heterogeneous relevant information chunks into actionable knowledge. To this end, the project focuses on the development of scalable semantic search, discovery and aggregation methods for large amounts of textual and multi-media data. Rather than just finding relevant documents, the focus will be on identifying semantically meaningful entities (such as people, groups, organizations, products, events, objects and scenes) and on relations between entities (such as experiences, attitudes, social roles, being-a-stakeholder). The project will emphasize *subjective information*—opinions, emotions, experiences, attitudes, voice characteristics—and *dynamics*: the changes in, e.g.,

entities, issues and stakeholders' positions tend to be even more important than the entities, issues and stakeholders' positions themselves. In addition, tools for truly semantic search have to be able to fuse information from disparate sources and modalities and to summarize the key strands.

The project's socio-economic key research question: *How can key developments relevant to an organization be identified, monitored, interpreted and aggregated to gain actionable insights?* This translates into the following key scientific research question: *How can methods for analyzing text and multi-media information deal with semantically rich multimedia information, in which dynamically changing factual and subjective aspects are intertwined key ingredients, and how can this be scaled to very large data volumes, be executed in near real-time, with outcomes aggregated to produce usable overviews?*

3. Objectives

Project's goal

Search technology has become a key component of our daily life. The ability to locate relevant documents has become a standard commodity. This project takes search beyond merely locating documents. Entities and events are seen as the most salient ingredients of textual and audiovisual content; they represent valuable knowledge ingredients that are heavily searched for, and conveniently serve as the linking pin between heterogeneous content sets. Semantic search is structured around entities, their properties, relations, developments as well as pertinent attitudes and emotions. The project takes on the challenge of building semantic search technology, with a special focus on issue management algorithms and tools, i.e., for identifying entities, issues, naming stakeholders, capturing attitudes and emotions, and monitoring relevant events and changes in highly dynamic environments.

Planning of all dimensions

P1 will be developed in four phases, with milestones at m9, m24, m36, m48. The initial phase will focus on requirements analysis, use case analysis, data set creation, system design and making available to all project partners of tooling available at the partners's sites as initial baselines. Later phases will consist of significant extensions to these baselines. Results typically emerge from analysis-designevaluate cycles for each milestone (in the form of papers, systems, data sets and the outcomes of user studies. Impact and valorization will be anchored in use cases and will mostly be achieved through the release (as open source or through licensing) of software. Dissemination is mainly pursued through national newspapers (NRC, Volkskrant), professional publications (Computable, Information professional, DIXIT) and public demonstrators. International embedding is realized through the organization of and participation in international competitions in information retrieval and language technology (TREC, CLEF, INEX, MediEval) and through coordination of and participation in ESF and EU FP7 programmes and international associations. Synergy is envisaged with other COMMIT projects that can benefit from P1's text-

based technology (P5, P6, P7, P23) or that complement or help facilitate it (P6, P20, P23). Within P1, researchers from the humanities and social sciences have a special and explicit role, with regards to methodological grounding and content analysis aspects of the project.

Results

The main scientific targets at the end of the project are 10 PhD theses. These will embody the knowledge to be developed to realize the scientific goal of building semantic search technology, with a special focus on issue management algorithms and tools. These theses will document solutions arrived at through problem-driven scientific research. They will be complemented with, and build on, software that be made available to project partners and beyond as open source or through licensing, thereby consolidating and sharing the knowledge and know-how created by P1.

Deliverable Impact and Valorization

The overall strategy to impact and valorization is through the release (either as open source or through licenses) of software that can be used or integrated by the project's technology providers as well as external partners. Concrete packages on which we will build are TIMBL, FROG, EARS, Fietstas, Lucene/SOLR, PF/Tijah, SHoUT, and Impala. Impact will be measured in terms of the number of partners and outside organizations taking up software deliverables, the number of licenses and the number of downloads of open sourced packages.

Deliverable Dissemination

Work concerning dissemination will focus on workshops aimed at professionals plus tutorials (beyond the MSc level) at the leading conferences. Furthermore, there will be outreach activities at national professional events, publications in the national press and in the professional press, and announcements in suitable LinkedIn groups. Existing scientific outreach activities conducted by P1 partners (with Religious Studies, Medical Anthropology, Television History, Communication Science and Medical History) will be brought under P1, using the services created by WORK PACKAGE5. Public demonstrators and/or videos will be jointly developed by each of Work package 1-4, 6-10 for dissemination purposes; we envisage zooming on three main demonstrators around information extraction, information retrieval and video search; public-facing demonstrators will be announced through appropriate social media channels.

International Imbedding

WP1-10 start from unique positions, either because of their approach, the tooling on which they build or the data to which they have access (or a combination of these factors). Uniquess will be guarded through repeated participations in worldwide evaluation efforts, including TREC, CLEF, MediaEval. Members of P1 will also be involved in the organization of these benchmarking activities. In addition, P1 project members participate in international cooperations through (and coordination's of) ESF and EU FP7 NoEs as well as through international associations; they will maintain or even expand their portfolio in the course of the COMMIT programme. Finally, in P1

uniquess is being created through collaborations between ICT specialists and leading researchers from the humanities and the social sciences.

Deliverable Synergy

The use of web service-based implementations of many of the tools facilitates rapid adoption within and outside the consortium. P1 envisages synergetic relations with the following COMMIT projects:

- With P5 – WP5 on recommender systems; the P1 project leader will act as formal promotor of the PhD candidate to be employed for this WP.
- With P6 – WP1 on using online multimedia as a training resource; With P6 – WP2 on video search; With P6 – WP6 on social interaction and content access modeling; With P6 – WP9 on large-scale metadata management.
- With P7 – WP2 on person-centric reasoning; we will (in WP2 and WP5) collaborate on language and information extraction technology to emerge from WP2 and WP5 and to facilitate their use in P7/WP2.
- With P20 – WP7 on scalable implementations of entity refinement; we will (in WP6) collaborate to acquire better scalable versions of the entity refinement tools on a map-reduce infrastructure. How would query plan rewriting techniques from PigLatin and real-time
- stream processing address our needs, and how can these be integrated in the application-type specific scheduling support for the DAS-4 supercomputer
- With P23 – WP2 on information access and distributed reasoning
- With P24 – WP1 functional data analysis.

4. Economic and social relevance

The core socio-economic problem addressed is that of semantic issue management, i.e., identifying key players and monitoring, anticipating and taking appropriate action on emerging trends or issues likely to affect an organization. The project considers this problem in multiple guises: (i) in the financial domain, where organizations wish to monitor a complex mixture of news, political data, financial data, user generated data, and opinion data; (ii) on social media, where knowledge-intensive companies wish to mine discussions for product and strategy development; and (iii) in multimedia settings, for high-tech partners that deal with large amounts of multimedia and multi-modal information.

Impact on Dutch economy: Surveys of the value of global brands (see, e.g., a November 2008 BusinessWeek/Interbrand survey) show that the amount of shareholder value attributed to companies' brand reputations is enormous, often in the billions. Issue management is meant to help protect brand reputations and thereby market cap. At least two dozen companies work in media analysis and co-creation consultancy in the Netherlands each; they stand to benefit from the project's tools. TrendLight will get new opportunities, being able to handle far larger accounts, with an estimated growth in turnover of at least 15%. Talking Trends is a joint venture

of UvA and TrendLight working on Dutch language tooling for PR and communication performance analysis, for whose ambitions this project is crucial. As news is becoming a commodity, the project is expected to help ANP maintain and expand its position by helping it turn into an information enrichment organization. IRN stands to gain by getting early access to text analysis tools for large volumes of web and social media data. Ilse Media, NIBG and De Persgroep will get new ways of providing economically viable access to their archives repositories, while Teezir, Textkernel and WCC will further innovate and refine their solutions to supporting such types of access. UvA spin-offs EUVision and ThirdSight will be able to exploit technology for challenging new search scenarios.

To address the socio-economic problem of semantic issue management, the project takes on challenges in *information retrieval* (to identify and track topics and stakeholders in dynamic data), *language technology* (to develop automatic frame analysis, to enable adaptive information extraction, to generate Dutch language summaries), *multimedia analysis* (to use pictorial cues for image/video search, to analyze body signals so as to enable extraction of emotion recognition).

Tremendous value has been created through the development of Dutch knowledge infrastructure for information retrieval and text and multimedia analysis, making The Netherlands the leading information retrieval country in Europe. However, robust automatic semantic issue management requires additional fundamental research and substantial advances beyond the state-of-the-art technology. Today's communication and (social) media analysis companies are not high tech and do not have the research facilities and multidisciplinary expertise required to realize these advances. Similarly, while the project's financial end-users do have extensive communication and PR departments, they lack communication researchers and research facilities. And the IT industry does not have the resources to develop the required technology themselves. Content-based video analysis of the type envisaged by the project requires a 2–5 year scope. For the first time ever in this area, the requested funding will bring together professional needs, academic resources, IT competencies, and market knowhow to speed up industrial development and uptake.

Turmoil in the financial markets indicates that in a rapidly changing socio-economic environment, responses and concerns from stakeholders (consumer, intelligence, financial, political, societal, advertising) can have dramatic impacts. Where actions can have a global impact in near real-time, there is a need for automated real-time monitoring tools that help identify the issues that affect an organization, that name the stakeholders that matter, and that track and predict emerging events. Further, in a competitive knowledge-rich society, tapping into an organization's brainpower may be a key to innovation, especially when skilled employees are hard to come by.

The web services developed in WP1–4 and professionalized in WP5 will continue to be hosted, maintained and supported by UvA, available for usage by all project partners; they form a key ingredient of UvA's long term research strategy in intelligent information access; moreover, the

services will be open standard, open sourced, allowing any partner or other interested party to integrate them locally. The technological deliverables of WP6–10 will be integrated by the profit partners involved in those WPs.

These concern (i) semantic issue management software components (for textual and multi-media data) released as open source and integrated by the profit partners; and (ii) issue management web services (for textual data) with client applications developed by the project, to be integrated by the profit partners.

5. Consortium

The consortium consists of leading researchers—a former Pionier grant holder as well as three current Vici grant holders and a Vidi grant holder. The WP leaders have complementary primary expertise areas, with sufficient overlap in secondary disciplines to ensure "a shared language" within the consortium:

Team, institute	*Primary discipline*	*Secondary discipline*	*WPs*
M. de Rijke, UvA	Information retrieval	Social media mining	WP2, WP5
T. Gevers, UvA	Content-based video analysis	Video retrieval	WP9
A. Smeulders, UvA	Multimedia retrieval	Computer vision	WP10
C. de Vreese, UvA	Communication science	Media analysis	WP4
A. van den Bosch, UvT	Memory-based learning	Information extraction	WP1
F. de Jong, EUR	Economic media analysis	Information extraction	WP3
A.P. de Vries, CWI	Information access	Intranet search	WP6
D. Hiemstra, UT	Web retrieval	Structured information	WP7
R. Ordelman, UT	Automatic speech recognition	Video retrieval	WP8

The WP leaders are well-known in their respective fields and have broad national and international networks from which project employees can be (and have been) recruited.

Roles: In addition to leading research teams, the consortium consists of a broad palette of partners, small as well as large, each with one of three roles: problem/owner, data provider or technology provider:

- Small partners (problem owner/use case): TrendLight (media analysis; will develop issue management use cases in the financial domain and in co-creation).
- Small partners (data provider): TrendLight.
- Small partners (technology provider): EUVision, Talking Trends, Teezir, Textkernel, ThirdSight.
- Large partners (problem owner/use case): ANP and De Persgroep (data enrichment to facilitate new business models, Internet Recherche Netwerk (monitoring open sources for radicalization).
- Large partner (data provider): ANP, Ilse Media, NIBG.

- Large partner (technology provider): Internet Recherche Netwerk, WCC.

The project's Work package can be grouped into three groups, one on the interface of information extraction and information retrieval (WP1–WP5), one on scalability and adaptability in semantic search engines (WP6–WP8) and on emotion in video (WP9–WP10). Within a group, the WP's involved will work on closely related problems, data and technology; and extensive sharing of, and collaborative work on, tools-in-progress, expertise and know-how is foreseen. Groups will collaborate in international competitions, as a Dutch COMMIT team (TREC Entity track, TREC Microblog track). To a large degree, the groups complement each other in terms of media type or modality; between groups, there will be an exchange of completed tools and components.

The consortium for P1 includes a work package based in a humanities department (WP1, Van den Bosch) and one based in a communication science department (WP4, De Vreese). Furthermore, several ICT-based work packages build on a long line of multidisciplinary work: WP2 and WP5 (ongoing collaborations with Religious Studies, Medical Anthropology, Television History, Communication Science, Medical History) and WP10 (ongoing collaborations with Communication Science).

6. Workplan

WP	*Partners*		*WP Nr*	*Partners*	
	Knowledge	*(Non)profit*		*Knowledge*	*(Non)profit*
WP1	UvT	ANP	WP6	CWI	Teezir, Textkernel
WP2	UvA	Talking Trends	WP7	UT	WCC
WP3	EUR	De Persgroep, Teezir	WP8	UT	NIBG
WP4	UvA	TrendLight	WP9	UvA	Ilse Media, ThirdSight
WP5	UvA	Politie Gelderland-Zuid	WP10	UvA	EUVision

The project will be developed in four phases, with milestones at m9, m24, m36 and m48. In the first phase (m1-9), WP5 will use existing toolkits and codes bases provided by UvA, UvT, CWI, and UT to set up baseline web services to be used by all partners; all other partners will focus on requirements (review papers, use cases, designs). At m18, m30, m42, WPs 1–4, 6–10 will deliver upgraded prototypes; WP5 will integrate those delivered by WP1–4 in the project's web services (thus reaching milestones 2, 3 and 4). The table below specifies the 5 core activity types of the project (left most column) together with types of content the project will work with. Each WP occurs in at least one unique activity/content cell, explaining its inclusion in the project.

Professionalization	WP5				
Methodological grounding	WP4				
Interpretation	WP2, 3	WP2	WP3	WP2,3	WP9, 10

Extraction	WP1		WP2		WP3	WP8	WP10
Retrieval	WP2, 6	WP2, 7	WP2			WP6, 8	WP9
	Edited content	*Web content*	*Social media*	*Panel data*	*Financial data*	*Speech*	*Video*

The overall aim of the project is to develop models, methodology, algorithms, and test and initiate the exploitation of issue management tools. At least 10 (non)profit partners will take up and integrate software components developed within the project within their production environments. Knowledge created will be documented in 9 PhD theses, at least 9 overview papers, 30 conference papers and 30 journal papers.

*M*ilestone 1 (m9): requirements (review papers, use cases, designs); baseline versions of software based on code bases brought in by partners; initial exploitation plan. Milestone 2 (m24): demonstrator software stage; initial exploitations by project partners, revision of exploitation plan. Milestone 3 (m36): golden demo stage; revision of exploitation plan. Milestone 4 (m48): final software deliverables and exploitation plan.

The risks: (R1) crawling hindered by overly restrictive server policies. (R2) use cases too specific or specialized. (R3) solutions do not scale well. (R4) insufficient training material for machine learning approaches. (R5) software developed by academic partners not sufficiently professional for uptake and integration. (R6) systems too inaccurate for deployment according to evaluation by end-users. (R7) quality or quantity of publications insufficient.

The remedies: (R1) enter negotiations with content owners. (R2) involve additional for-profit partners at worktables. (R3) reduce functionality; explore service-based asynchronous solutions. (R4) use combinations of active learning, bootstrapping, semi-automatically labeled data. (R5) software developed in WP1–4 will be professionalized in WP5; software in WP6–10 will be jointly developed with profit partners. (R6) more target oriented system, more focused datasets. (R7) shift emphasis from engineering to science.

References

[1] K. Balog et al., A Language Modeling Framework for Expert Finding. *Inf. Proc. Manag*, 45:1-19, 2009.
[2] J. Bliter, R. McDonald, F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP 2006*, 2006.
[3] A. van den Bosch, W. Daelemans. *Memory-Based Language Processing*. Cambridge University Press, 2005.
[4] J.G. Shanahan et al., editors. *Computing Attitude and Affect in Text*. Springer, 2006.
[5] D.R. Radev et al., Centroid-based summarization of multiple documents. *Inf. Proc. Manag.*, 40: 919-938, 2004.
[6] R. Barzilay, M. Lapate. Modeling Local Coherence: An Entity-based Approach, *Comp. Ling., 34*, 1-34, 2008.
[7] E. Krahmer, S. van Erk, A. Verleg. Graph-based Generation of Referring Expressions, *Comp. Ling., 29,* 53-72, 2003.
[8] C.H. de Vreese. News framing: Theory and typology. *Information Design Journal + Document Design*, 13: 48-59, 2005.
[9] P. Sinha, R. Jain, Classification and annotation of digital photos using optical context data, In: *CIVR 2008*, 2008.
[10] K. van de Sande et al., Evaluation of Color Descriptors for Object and Scene Recognition. In: *CVPR 2008*, 2008.
[11] A. Chaudhury. How Brand Reputation Affects the Advertising-Brand Equity Link. *J. Advertising Research* 42:33-43, 2002.
[12] P. D'Angelo. News framing as a multi-paradigmatic research program. *Journal of Communication*, 52:870–888, 2002.
[13] J. Turmo, A. Ageno, N. Català. Adaptive information extraction. *ACM Comput. Surv*. 38(2), 2006.
[14] R. Soricut, D. Marcu. Towards developing generation algorithms for text-to-text applications. In: *ACL 2005*, 2005.
[15] M.J. Halvey, M.T. Keane. Analysis of online video search and sharing. In *HT '07*. 2007.
[16] A.F. Smeaton, S. Rothwell, Biometric responses to music-rich segments in films: the CDVPlex, In *CBMI 2009*, 2009.
[17] A. Mahajan et al.. Mining Financial News for Major Events and Their Impacts on the Market. In *WI-IAT* 2008, 2008.
[18] G. Mishne, M. de Rijke. A Study of Blog Search. In: *ECIR 2006*, 2006.
[19] D. Newman et al., Analyzing Entities and Topics in News Articls Using Statistical Topic Models. In: *ISI 2006*, 2006.
[20] O. Etzioni, M. Banko, S. Soderland, D.S. Weld. Open information extraction from the web. *Comm. ACM* 51: 68-74, 2008.
[21] C. Mota, R.Grishman, Is this NE tagger getting old? In: *LREC 2008*, 2008.
[22] B. Favre et al., Punctuating Speech for Information Extraction. In: *ICASSP 2008*, 2008.
[23] A. Culotta et al. Corrective feedback and persistent learning for information extraction. *Art. Int.* 170: 1101-1122, 2006
[24] J. Wang et al., Brain State Decoding for Rapid Image Retrieval. In ACM MM *2009*, 2009.

WORKPACKAGES

Project number P01	
WP title acronym	Adaptive Information Extraction over Time (ADNEXT)
WP leader	Antal van den Bosch, Universiteit van Tilburg

Objectives:

To develop trainable, adaptable Dutch language information extraction technology for named entity recognition, event detection, and time identification, as key components in the P1 target of Semantic Search. The technology has a broad coverage "default" mode and retrains dynamically to new domains upon being confronted with new (clusters of) news or user-generated data.

So far, most information extraction technologies have been developed with broad coverage in mind. However, in dynamic domains such as news, new texts contain many new events, and introduce relatively many new entities. Adaptation to high degrees of novelty in text is a prerequisite for successful deployment of information extraction technologies. With adaptation comes a need for integration with active time management (and the detection of time expressions), and dynamic semantic networks of world knowledge such as Wikipedia. Application: Knowledge enrichment (in news and user generated data)

Task 1.1: dynamic entity recognition (UvT, ANP). Task 1.2: dynamic time expression detection and normalization (UvT, ANP). Task 1.3: dynamic event detection (UvT). Task 1.4: integration (UvT, ANP).

Project number P01	
WP title acronym	Stakeholder Tracking (ST)
WP leader	Maarten de Rijke, Universiteit van Amsterdam

Objectives

Objective: To develop, implement and test algorithms for identifying relevant topics (for a given problem owner), for identifying the stakeholders that entertain a position on those topics in news data and in user generated content, and for signalling significant changes in topics of stakeholders' positions.

Background: Topic detection and tracking has been an active research area for many years, but restricted to edited news content and to facts rather than entities. Robust mining of associations between topics and entities has mostly been limited to expertise in the case of academics or other knowledge workers, but needs to be extended to "perspectives"—moreover, the associations need to be explained and equipped with brief supporting evidence. First attempts at automatic aggregation of perspectives were recently evaluated at the Text Analytics Conference. Large scale analyses of positions have so far been limited to static positions only.

Application: The core application is issue management, in both edited and user generated content.

Description of work: UvA will develop models, algorithms and prototypes to be professionalized in WP5; Talking Trends will provide labelled data and aid in specifying evaluation criteria. Task 2.1: modelling themes and their dynamics using language models and (dynamic) topic models (UvA, Talking Trends). Task 2.2: Identifying stakeholders and characterizing their positions using entity recognition, relation extraction and association mining (UvA, Talking Trends). Task 2.3: aggregation of static stakeholders' positions using graph-based methods (UvA). Task 2.4: Identifying and predicting changes in stakeholders' positions (UvA, Talking Trends).

Project number P01	
WP title acronym	Mining Economic Entities (MEE)
WP leader	Franciska de Jong, Erasmus Universiteit Rotterdam

Objectives

Objectives: To design and develop entity detection for the domain of economic news and to exploit it for the realization of trend mining tools for economic news content.

Background: The envisaged approach is to develop a framework that can accommodate at least the following three parameters: media format (text – multimedia), data type (full text only - numerical economic indicators), and chronological perspective (contemporary content – historical archives). The tools should be of interest to news channels and platforms, and to digitized historical news archives. For the former, the tools should support facetted semantic content interpretation across media types and integration with numerical data types that are in use as indicator of economic trends. For the latter, the tools should deliver pertinent content scholars in economics and history and help them refine their models.

Application: 1. Economic barometer based on the integration of textual and numerical data. 2. Timeline visualization techniques for historical news archives

Description of work: There is an intricate collaboration with the industrial participants, because they are essential in providing the data and user requirements for the analysis and cross-media mining functionality as basis for the development of new content services for their domain (Persgroep), and for the experimentation with entity detection and visualization of trends (Teezir). For EUR there is a dual added value in setting up collaboration between computational economics and experts in entity detection. First of all because it will stimulate the innovative coupling of data mining technology to text and media mining, and secondly because models for the generation and transfer of economic insights can now be checked against multifaceted historical data sets. Task 3.1 Development of training corpora (EUR/Persgroep), Task 3.2 Identification relevant economic indicators and model for integration with text mining (EUR), Task 3.3 Tool development and usability testing (EUR/Teezir /Persgroep)

Project number P01	
WP title acronym	Computational Framing (CF)
WP leader	Claes de Vreese, Universiteit van Amsterdam

Objectives

Objective: To develop and validate tools for automatically linking frames in communication with frames in public opinion in the area of economics and politics. To investigate the dynamics and impact of frames in difference channels.

Background: Frames are ways in which issues are presented, emphasizing the salience of certain aspects and downplaying other aspects. Our current research of how frames affect opinions and evaluations is based on static research designs and there is a need for more detailed and dynamic understanding of evaluations. In the economic arena, knowledge of sentiments and emotions is pertinent for political elites and businesses that have to operate in a global and rapidly changing market where consumer responses can make or break a business.

Application: The core application is issue management, especially in economics (relation to WP2 and WP3) and politics (relation to WP2).

Description of work: The industrial partners (through TL) will provide use cases, data and inform the experimental task settings; UvA combines its strengths in communication and information technology. Task 4.1: query modeling for frame finding (UvA, TL). Task 4.2: linking extracted perspectives with public opinion data on economic perception (UvA, TL). Task 4.3: develop a dynamic model of changes in the politico-economic information environment and party evaluations and support that depends on economic as well as political reporting (UvA, TL).

Project number P01	
WP title acronym	Software Development (SD)
WP leader	Maarten de Rijke, Universiteit van Amsterdam

Objectives

Objectives: To consolidate and validate modules developed in WP1-WP4 and to integrate these into a set of robust web services that support issue management "in the large" (in news, politics, user generated data) and "in the small" (on a specific discussion platform). To launch the web services early on in the project and to incrementally update them with the project's results. To aid project partners in preparing modules for integration in the project's web services. To aid partners in writing client applications that use the project's web services.

Background: Web services have become a popular way of providing data, compute or knowledge intensive computational facilities across the web. A particularly relevant example is Reuter's OpenCalais service that provides an entity recognition service through a standard API. UvA's Fietstas service is another example for the Dutch language. The project's services will go beyond this by (additionally) providing stakeholder, perspective, and frame finding services as well as a dynamic summarization service.

Application: The core applications are issue management and knowledge enrichment.

Description of work: Task 5.1: baseline web service with data gathering, search, simple annotation and term clouding functionality; uptime at least 90%; at least 10 online news sources, 1000 blog-like sources, all relevant parliamentary sources ("handelingen," "kamervragen," "moties"); able to annotate at least 10,000 items per day (UvA, IRN). Task 5.2: extend sources, uptime at least 95%, extended annotations (frames, dynamic topics, stakeholders, extractive summaries) (UvA, Talking Trends). Task 5.3: improved uptime, added functionality (dynamic stakeholders, events); golden demos selected from the project's applications (UvA, IRN). Task 5.4: uptime at least 99%, added functionality (generative summaries); preparing for post-project handover (UvA, IRN).

Project number P01	
WP title acronym	Context-Aware Entity Refinement (CAER)
WP leader	Arjen P. de Vries, Centrum Wiskunde Informatica

Objectives

Objective: To improve entity ranking by shifting entity detection from indexing time to query time. To create an open source solution for entity ranking in speech sources (with WP8).

Background: Existing entity ranking solutions depend on a pre-defined (and usually short) list of entities detected at indexing time (e.g., person, location, organization, and other). The resulting semantic annotation (or tagging) of the indexed documents could in principle support better results for information needs involving people, locations or organizations. However, the general nature of these entity types is often a poor match with the information need at hand. One approach to handling this problem is to predict what information needs are to be supported, and build task-dependent parsers (taggers) accordingly. This WP also investigates a novel alternative called *entity refinement*: shift a (partial) entity detection step from indexing to search. The advantage would be that the tagging can be specialized to the information need at hand, using all the additional information known at query time. The challenge here is to tackle two problems: run-time efficiency and availability of training data.

Application: Entity ranking on speech sources, where the most representative features for entity detection are with high likelihood out-of-vocabulary words.

Description of work: Task 6.1: incrementally trained task-dependent taggers (Textkernel); Task 6.2: entity refinement (CWI); Task 6.3: entity ranking on speech sources; apply entity refinement to SHoUT automatic speech recognition (CWI, with WP8); Task 6.4: dynamic aspect extraction; Task 6.5 evaluation; participation in TREC web entity entity ranking (CWI, Teezir); Task 6.6 scalability (CWI, with P20).

Project number P01	
WP title acronym	Deep Web Entity Retrieval (D-WER)
WP leader	Djoerd Hiemstra, Universiteit Twente

Objectives

Objectives: To open up deep web sources for entity retrieval by identifying the entity types a web source provides, by allowing natural (text search) access to structured data, and by combining entities from diverse deep web sources. WP 7 aims to provide possibilities to share and search information, without the need for the service provider (WCC) to crawl all data.

Background: Recent research in entity retrieval has resulted in effective entity ranking if the task is well-defined as in expert search, or if the data is well-organized as, e.g., in Wikipedia. Well-defined and well-organized data is increasingly available on the web—the prime example being the deep web. The deep web is a large part of the web that cannot be accessed by crawlers: mostly dynamic web pages that are returned in response to a web form. The objectives are met in four steps. *Personal information sharing* evaluates a prototype content management and search system based on WCC's matching system ELISE at UT. *Deep web entity probing* identifies what types of entity a deep web service provides by probing queries. A challenge is to identify exactly what entity types a web service specializes in, but general types are of interest as well, for instance "persons." *Database natural abstraction layers* opens up a (deep) web service by returning dynamic pages based on text queries or natural language questions, combining closed-domain question answering approaches with open-domain approaches, so as to identify question patterns that a deep web source may answer and automatically translate questions to web forms. *Entity search aggregation* will combine deep web information from several sources in a unified search result. This work will focus on the use of standards like OpenSearch and on extensions to support deep web entity retrieval. In the absence of standardized results we investigate wrapper induction techniques for information extraction.

Application: personal information management based on WCC's ELISE system; distributed information retrieval

Description of work: Task 7.1: personal information sharing (UT, WCC); Task 7.2: deep web entity probing (UT,WCC); Task 7.3: database natural abstraction layers (UT, WCC); Task 7.4: entity search aggregation (UT, WCC).

Project number P01	
WP title acronym	Spoken Entities (SpEnt)
WP leader	Roeland Ordelmans, Universiteit Twente

Objectives

Objective: To locate and confine spoken entity references (e.g., named entities but in a broader sense for example also quotes) in very large quantities of unstructured audiovisual content. To accumulate structural, semantic and supra-segmental information that can be used for identifying spoken entities in context. To optimize speech and audio tools for the task characteristics (focus on entities, large data). To identify strategies for audiovisual spoken entity exploration and cross-linking.

Background: Locating and confining spoken entities in context is challenging. Being able to identify these important information markers in AV content, allows fine-grained, facetted access that in turn improves the exploitability of rich AV content significantly. In very large and expanding data sets, named-entity occurrences are hard to predict and tend to be out-of-vocabulary. As named-entities have a foreign or historic origin, their pronunciation may diverge from the pronunciation predicted by pronunciation models for the target language. Connecting progress in spoken term detection with trends in Information Retrieval (e.g., Entity Ranking) is an important next step. By exploiting the potential of various types of information that are available in the audio track, spoken entity location could go beyond pinpointing words or word groups, allowing a broader interpretation of entities including speakers, quotes or affect categories.

Application: search in multimedia archives

Description of work: In collaboration with Sound and Vision and PCM, a demonstrator will be developed that on the basis of large amounts of textual news and AV data (radio and television) shows the added value of multimedia entity exploitation. Task 8.1 Baseline audio and speech processing development; Task 8.2 Entity occurrence prediction from metadata and context; Task 8.3 Combined ASR Search modeling (1-best, term spotting, lattices, hybrid approaches, OOV recovery); Task 8.4 Entities in context; Task 8.5 Integration.

Project number P01	
WP title acronym	Semantic Video Search on Mobile and Web platforms (SVSMW)
WP leader	Theo Gevers, Universiteit van Amsterdam

Objectives

Objectives: To provide image/video access as an information service and entertainment on web and mobile platforms. Large digital archives are currently accessed by meta data. Integrating efficient content-based access will enlarge the usability of the archives used by different communities.

Background: This WP will focus on semantic search and services to support event, news and entertainment analysis on the Internet and mobile phones especially for video platforms such as zie.nl owned by Ilse Media. Although Ilse Media currently delivers a platform for video storage, indexing and search are limited to key words. Content-based search is urgently needed. Ilse Media lacks the knowledge or expertise on content-based video access. Large scale datasets will be made accessible by Ilse Media and a beta-site will be issued to gain experience in "semantic video search in the wild". This provides this WP with documented datasets, user-driven questioning, and content interaction. Semantic search has so far been targeted at a single information source; this WP combines multiple modalities to go beyond standard video retrieval. The modules developed by UvA are integrated in a demo system by Ilse Media and made available for public use (mobile and desktop), increasing the societal and economic value of semantic video search.

Application: mobile video search issuing more requirements on community and sharing of video.

Description of work: Task 9.1: Based on user study, valuable concepts are identified for different application areas (UvA, Ilse Media). Task 9.2: techniques and software are developed for large scale processing of features and classifiers for concept and event recognition in videos suited for web (zie.nl) and mobile phone platforms (UvA, Ilse Media). Task 9.3: techniques are developed for image quality analysis and shot representation for video processing and summarization (UvA). Task 9.4: information fusion is developed for video search based simultaneously on text, annotations, speech and pictorial modalities (UvA, Ilse Media).

Project number P01	
WP title acronym	And Looks-like Amazing News (ALAN)
WP leader	Arnold Smeulders

Objectives

Objectives: The saying among directors is “television is emotion”. In this WP, we start from that wisdom for the purpose of making available emotion-differentiated concepts when gathering information from large collections of video. For this purpose emotion may be divided into: “energizing, cheering, saddening, depressing, distressing, comforting”, and other states. For the moment we will select 10 initial states.

Background: This WP aims to establish algorithms that have a good predictive power on each of these aroused states in the viewer. The aim is to predict the state from the responses of audio-visual detectors yet to be learned from example videos. In contrast to the standard method of training a concept detector in visual search, we will not use explicit training from example images but rather use time based physiologic signals acquired from persons when watching television: heart-rate, reduced EEG-signals, sweat, and eye or body movements. The television to watch will both be directed as well as undirected footage (commercials, motion pictures, news, YouTube) to quantify the influence of directing.

Application: Advertising.

Description of work: Sophisticated multi-variate variance decomposition by functional data analysis will be required to learn which physiological signals relate to what emotional state visual detectors. Inspiration is derived from Alan Smeaton. Sennay Ghebreab, Victor Lamme en Steven Scholte are possibly interested to predict the emotional value of movies. See also the Kuleshov effect (Cees Snoek). Advertising agent EQ brands is prepared to make available their data (commercials, storyboards). Some fMRI data are there already.

DELIVERABLES

Number of important journal paper
30

Number of important conference contributions
30

Products

1. Use cases and design event

   The two involved parties will define use cases based on a general setting of an online, automatic, real-time news monitoring service that detects significant events in newswire, and optionally in social media. As archival data is available, use cases can both be off-line and online. Off-line use cases focus on the detection of events of a certain type through time, with the goal of detecting the same type of events later, possibly even before they happen (e.g. acts of protest such as strikes are often preceded by other actions signalling social unrest). On-line use cases may involve a strict newswire context, where the focus is on the correct identification and time stamping of novel events (in contrast with news items that report on continuations of events), or may involve a mixed context where online news is tracked both in the newswire context and in social media such as Twitter and Facebook. Proper evaluation methods will be aligned with the various use cases, and annotation tasks will be identified.

   - Work package 5, Year 2012

2. Report on use case analysis of issue management

   The parties involved will define use cases based on the general setting of online, real-time issue monitoring in social media. Interviews will be conducted with customers of Talking Trends and with customers of Llorente Cuenca. In addition, issue management around political data streams will be analyzed in collaboration with stakeholders in the Political Mashup infrastructure project running at UvA (FNWI). Historical data will be examined and analyzed in collaboration with UvA FMG partners (involved in WP4). The focus will be on the complementary nature of multiple data streams, on information diffusion and on the emergence of unknown stakeholders.

   - WP 2, Year 2011

3. System design issue management

   The parties involved will conduct interviews with Talking Trend customers in issue management and co-creation so as to identify core ingredients of issue management systems. Through contacts within the CLEF Cross-Language Evaluation Forum and the PROMISE Network of Excellence, initial findings will be subjected to feedback of international partners active in online reputation management and co-creation (esp. Llorente Cuenca). Use cases will concern

multiple tasks, including source discovery and tracking, use edited news data (made available through WP1) as well dynamic data streams and user generated content (Twitter, Hyves and Facebook). The deliverable will also inform evaluation methods to be deployed at a later stage for assessing the output of issue management systems.

- WP 3, Year 2011

4. Text corpus tagged for economic entities

   In order to develop text mining techniques to capture the economic sentiment, the construction of a corpus is need that is manually of semi-automatically annotated with tags for those linguistic entities that contribute to the economic sentiment, such as names of companies, names of events (such as take-overs, interest changes, stock price development, etc.) and terms that express the trust among politicians, policymakers and the general audience in the economy. Also certain elements marking the rhetorical structure of a text could affect the weights of certain sentiment markers, so they need to be tagged as well. The corpus will play a role in the training of models and in the evaluation of text mining performance. Before the actual tagging can take place, the required design and size of the corpus needs to be investigated.

   - WP 4, Year 2012

5. Use cases and design automatic framing

   The two parties involved (UvA (FMG) and TrendLight) will survey recent case studies, of either commercial or academic interest. Annotations will be analyzed and methodological differences and commonalities between academic and commercial ways of working will be identified. Use cases will be examined from diverse societal sectors: political, economic, medical, financial, cultural. The balance between relatively shallow but rapid

   - WP 5, Year 2012

Software

1. Body signals measurement

   UvA will complement equipment for measurement of body signals (heart rate, light EEG, sweat, motion, visual) plus subsequent data analysis synchronized to film.

   - WP 10, Year 2012

2. Top-k techniques in PF/Tijah

   The existing PF/Tijah open source text search developed at the University of Twente, the University of Tuebingen and the University of Konstanz will be extended with top-k techniques to rank query answers, with a special focus on ranking and retrieving entities.

   - Work package 6, Year 2012

3. Image feature extraction and classification

   UvA and Ilse Media will release as software a set of techniques developed for large-sclae

processing of features and classifiers for concept and event recognition in videos suited for web and mobile phone platforms.
- Work package 9, Year 2013

4. Time expression module
   Based on annotated data created in COMMIT, UvT will develop a time expression module for detecting absolute and relative time expressions for correct time stamping of events.
   - WP 1, Year 2013

5. Economic entity miner and visualizer
   The three partners involved, EUR, Persgroep, Teezir) will complete the iterative development of a mining and visualization environment for economic entities. The toolkit will couple data mining technology with text and media mining.
   - WP 3, Year 2014

6. SHoUT open source speech recognition toolkit
   UT and NIBG will complete iterative releases of Twente's large vocabulary continuous speech recognition toolkit, SHoUT. The final COMMIT release will integrate spoken entity detection enhancements as required by broadcast professionals, documentalists and television researchers.
   - WP 8, Year 2014

7. Automatic deep web entity retrieval
   UT and WCC will complete the iterative development of a deep web entity search engine, that comprises of automated probing, wrapper and ranking modules.
   - WP 7 YP 2014

8. Stakeholder tracking system
   Based on technical insights gained from participations in the TREC Entity and Microblog tracks, UvA and Talking Trends will build on iterative releases of Fietstas and EARS (WP5), integrating dynamic topic modeling and aggregation components with impact predictors, and release a stakeholder tracking system.
   - WP 2, Year 2015

User studies

1. User study spoken entities
   A requirements analysis will be conducted involving documentalists, broadcast professionals, researchers at Beeld en Geluid. Possible use scenarios involving spoken entities will be examined and feed into the creation of an evaluation benchmark, both for project internal and international competition purposes.

- WP 8, Year 2011

2. User study WCC software
   As a first step towards deep web entity retrieval, this personal information sharing step evaluates a prototype content management and search system based on WCC's matching system ELISE at UT, using students as testers and testees.
   - WP 7, Year 2011

3. User study emotion
   A user study will be conducted, involving both UvA and EUVision, on a group of people, aimed at assessing the effectiveness of emotion detection equipment based on physiological features.
   - WP 10, Year 2013

4. User study entity refinement
   CWI, Textkernel and Teezir will jointly carry out a user study aimed at examining the effectiveness of ad-hoc query-dependent tagging. For the study bot textual and web data will be used.
   WP 6, Year 2013

5. User study economic mining of historical content
   A user study will be conducted jointly by EUR and De Persgroep on mining historical content for economic entities and facts. Data will be provided by De Persgroep. System development and tuning will be based on annotated data made available within WP3. Multiple types of users will be considered and contrasted.
   - WP, 3 Year 2014

6. Field trial semantic video search
   Description A field trial will be conducted aimed at examining "semantic video search in the wild." UvA and Ilse Media will integrate semantic and location based modules in a demo system that will be made available for public use, in a mobile and desktop setting.
   - WP 9, Year 2015

7. User study emotion prediction
   Description UvA and EUVision will evaluate methods to define new effective emotions from the study of invariant patterns on a group of people.
   - WP 10, Year 2015

Other results

1. PhD thesis

PhD thesis on event mining.

- WP 1, Year 2015

2. PhD thesis

   PhD thesis on stakeholder tracking.

   - WP 2, Year 2015

3. PhD thesis

   PhD thesis on mining the impact of news on social media.

   - WP 2, Year 2012

4. PhD thesis

   PhD thesis on mining economic entities.

   - WP 3, Year 2015

5. PhD thesis

   PhD thesis on computational framing.

   - WP 4, Year 2015

6. PhD thesis

   PhD thesis on context-aware entity refinement.

   - WP 6, Year 2015

7. PhD thesis

   PhD thesis on deep wen entity retrieval.

   - WP 7, Year 2015

8. PhD thesis

   PhD thesis on retrieving spoken entities.

   - WP 8

9. PhD thesis

   PhD thesis on semantic video search.

   WP 9

10. PhD thesis

    PhD thesis on emotion detection in video from physiological features.