# Pushing the Quality Level in Networked News Business: Semantic-Based Content Retrieval and Composition in International News Publishing

*Markus W. Schranz*[1,2]

[1]Distributed Systems Group, Institute for Information Systems, Vienna University of Technology
Argentinierstrasse 8, 1040 Vienna, Austria
[2]Research & Development, pressetext Nachrichtenagentur GmbH
Josefstädter Strasse 44, 1080 Vienna, Austria
e-mail: schranz@pressetext.at; schranz@infosys.tuwien.ac.at

## Abstract

Electronic publishing exploits numerous possibilities to present or exchange information and to communicate via most current media like the Internet. By utilizing modern Web technologies like Web Services, loosely coupled services, and peer-to-peer networks we describe the integration of an intelligent business news presentation and distribution network. Employing semantics technologies enables the coupling of multinational and multilingual business news data on a scaleable international level and thus introduce a service quality that is not achieved by alternative technologies in the news distribution area so far. Architecturally, we identified the loosely coupling of existing services as the most feasible way to address multinational and multilingual news presentation and distribution networks. Furthermore we semantically enrich multinational news contents by relating them using AI techniques like the Vector Space Model. Summarizing our experiences we describe the technical integration of semantics and communication technologies in order to create a modern international news network.

**Keywords:** news management; semantic-based content coupling; news distribution

## 1 Introduction

Electronic publishing on the Internet has introduced a wide application range within the information management field, including the digitization of ancient archives, the manageability and availability of large amounts of data or the visualization and provision of contents to the open public [12,15]. Modern intelligent approaches have enriched the field of online information management with the research results of information retrieval methodologies, eLearning, theories of infocracy, security, privacy, semantic relations and metadata management towards the Semantic Web [2].

A specific application area, namely online news management and distribution has profited significantly by research results in the areas of distributed systems, information management and retrieval and even semantic-based techniques. Renowned news providers use research results to offer content clustering and links to semantically related news to external supplementary contents as introduced by for example the BBC Newstracker, and international search machines like Google or Yahoo offer dedicated services to observe most recent news structured in e.g. related topic clusters.

Recent research work within the EU-funded project NEDINE [10] focuses on the creation and the conception of an intelligent news publishing and distribution network consisting of existing local news agencies that use modern Internet and distributed computing technologies to build up a new kind of multinational and multilingual news distribution service. Research results [1,12,16,20] from the area of network communication have been involved to create a scaleable loosely coupled service architecture and artificial intelligence is utilized to identify the most relevant related news articles within the investigated multinational Nedine service. For semantic relationship management, the Vector Space Model [13, 14, 17] has been thoroughly investigated and applied to provide easy access to related and most relevant business news.

This paper is focusing on the combination and the integration of the research results in order to join modern Internet technologies such as Web Services [6] and peer-to-peer architectures [1] with semantic-based content coupling and document composition to create a both scalable and qualitatively high-level news exchange and distribution network. We show, how semantic technologies and loosely coupled services are able to enrich business news networks far beyond the possibilities that local services and recent implementations could offer. A brief introduction to the problem area and the specifics of the application domain news management and distribution is presented in section 2. The consecutive section 3 outlines our methodological approach towards an

adequate service architecture for scaleable international news distribution whereas section 4 focuses on the semantic-based content enrichment by coupling highly relevant related news articles. Section 5 discusses results within an experience report based on the application of the earlier mentioned concepts in the EU project Nedine. A brief summary and an outlook to further development and practical use conclude the paper.

## 2 Challenges for International News Management & Distribution

International business today is highly segmented and widely unrecognized beyond national and language borders. Business news mostly bear national relevance but hold the potential to spread cooperation opportunities and business changes towards an improved integration of international businesses and multinational and multilingual information flow. The application domain we are focusing on has built a business on top of the technological features and basic services available on modern infrastructures such as the Internet: Hundreds of PR-companies and news agencies all over the world have utilized modern synchronous and asynchronous Internet technologies such as web-based information access front-ends for business news archives, email-based mass news distribution or content provision for next generation user devices. A media industry has emerged that uses complex and technologically challenging Internet services to create, aggregate, exchange, publish and distribute current business news.

International search engines have created dedicated news services to observe hundreds of pre-selected news providers to be able to offer a Meta-search engines on News contents. Leading examples are available at e.g. http://news.google.com/ or http://news.yahoo.com/. Renowned news providers like the BBC at http://news.bbc.co.uk/ offer themselves links to external sources to provide a comprehensive overview on a specific topic. With means of modern technologies news items can be clustered according to various criteria such as geographical origin, content language, content topics, etc.

Due to the importance of local competences, international high quality news management has been restricted to very few news agencies and a small set of privileged agencies that retrieve the often regionally very sparse information. In contrast, regional and national news providers and news agencies are empowered with strong local competence but are not able to reach a critical mass of readers or subscribers to build up a high quality news service. Existing services are based on modern Internet technologies like Web application servers holding the business logic in the services middleware [4] and maintaining a local network of content providing editors and commercial customers as well as thousands of subscribers.

In order to reach beyond national and language borders, scaleable technologies and improved content relation features are necessary to empower an integrated distributed network of competent local news agencies. Technologically, recent implemented services lack homogenous implementation models, data structures and communication protocols. With modern Internet technologies like Web Services for the information exchange and P2P architectures to manage a scalable integration of several local service providers a meta-network has been designed within the scope of this research work.

The implemented network discussed within this paper is targeting at the integration of such businesses, thus allowing news agencies of different countries to share contents and exchange business news towards an integrated network for news aggregation, creation and dissemination. Aside from the obvious business benefits of such service integration, there are necessary steps to be taken to technically and organizationally bring the services and the systems of the existing news agencies to an integrated network. Beyond this, further development shall attract additional agency partners to join the network in order to:

(1) access relevant business news at an international level;
(2) offer a distribution and dissemination interface for their customers that provide news to the network.

As a modern extension to existing multinational news networks we propose the utilization of semantic-based algorithms to automatically relate most relevant business news to each other. Modern research approaches such as the Vector Space Model and a specific adaptation for the use in business news management are designed to be used in both our research work and in consequence within a real business implementation (see sections 4).

The combination of modern Internet technologies and the application of semantic-based content interrelationship management has been a challenging research area with promising results for early integration in a business scenario as described in section 5.

# 3 A Proposed Service Architecture for Scaleable Networking

Internet technologies to be applied in the area of business news distribution involve technical features to manage scalability and performance in mass information provision (dozens of millions of page impressions) and mass distribution (millions of electronic mails sent daily). Since scalability is managed by the concept of integrating the local strength of existing services into a powerful network, modern networking features and capabilities are within the focus of the technical architecture.

## 3.1 Loosely Coupled Services: P2P Networking

Comparing the two approaches centralized news management server vs. peer-to-peer networking, we have identified several arguments that are in favour resp. in contradiction to the employment of the discussed network topology. Reasons to design a multilingual and multinational network based on a centralized news brokerage service include:

(1) having a single point of maintenance;
(2) clarity and simplicity of the infrastructure;
(3) only one traffic channel from agencies to the service;
(4) no additional infrastructure required for network partners.

On the other hand, our research group had to consider, that a centralized architecture also represents a single point of failure and an external source of maintenance and service costs. Apart from technical and administrative obstacles, business partners have influenced the research work with specific preconditions:

(1) no centralized storage of secret business data like customers, subscribers addresses or accounts;
(2) no duplicate storage of business news articles, endangering the copyright situation of the owner.

The proposed solution is to attach identical peer software components to the existing services of each network partner which allows each partner to locally register provided services according to a predefined schema of internationally offered features. Modern communication technologies such as Web Services via SOAP [5] and WSDL [18] are used to create the decentralized system, interconnecting all participating news agencies. The core architectural concepts handle data management, the publishing interface, and the interface connections. The interfaces between the communication peers and the interfaces between a peer and an existing local system are defined in detail by Web Services descriptions.

The peer concept to loosely couple the services was chosen since technically each potential partner can be equipped easily with the same piece of software. The network communication can be handled within the peer logic, only the interface between the peer and the existing service needs to be implemented by the joining partner. All activities within the multinational and multilingual network are triggered by the actions of one of the participating partners. The communication types within the system include Service registration, information upload, multinational news distribution requests and news enrichment requests.

## 3.2 Data Formats for the Content Exchange

Although most significant information in the news area is stored in traditional text files, the information management in the news application area has been modernized based on the developments in document format standards and XML. The International Press and Telecommunication Council IPTC [7] has been developing news formats and standards to capture data and meta-information on news, following the specific needs and requirements of the multimedia news industry. Most recently, IPTC's activities have primarily focused on developing and publishing Industry Standards for the interchange of news data, namely NITF (News Industry Text Format, current version 3.2), and NewsML [11].

NewsML can be applied at all stages in the (electronic) news lifecycle. It would be used in and between editorial systems, between news agencies and their customers, between publishers and news aggregators, and between news service providers and end users. Because it is intended for use in electronic production, delivery and archiving it does not include specific provision for traditional paper-based publishing, though formats intended for this purpose - such as the News Industry Text Format (NITF) can be accommodated. Multimedia content types such as image formats, audio- and video files are integrated with appropriate markup and description elements in the NewsML language. Additionally, the IPTC has created a standardized catalogue of news

categories which can be used to structure contents according to a news specific ontology (IPTC subject codes [19]). Semantic technology can be used to automatically identify news categories for individual news items (see also section 4.1).

We have selected NewsML as the appropriate content format for the international integration following a thorough research and evaluation period. News agency partners will have to adopt their local services to the current standard, thus providing interfaces that support NewsML contents for the Nedine exchange protocols.

## 3.3 Communication with the Peer Network

In order to integrate existing local services into a multinational network, communication protocols had to be defined and established, ideally following state-of-the-art models and research standards. We have chosen Web Services as the most appropriate and modern exchange mechanism, built on top of a Peer-to-Peer architecture to integrate the national services into a multinational network.

The communication between the peer software and the existing services follows the Web Service definitions in the Nedine WSDL documentation. The communication includes the upload of local news data, the registration of local services to the network, the distribution initiation and parameters as well as the regular polling of news data to be distributed on behalf of the network partners. The aim of the network architecting process was to keep the WSDL as small as possible to attract members to the technical network and enforce the construction of a Pan-European cooperative service of locally strong news agencies.

The utilized SOAP queries can be roughly describes as follows:

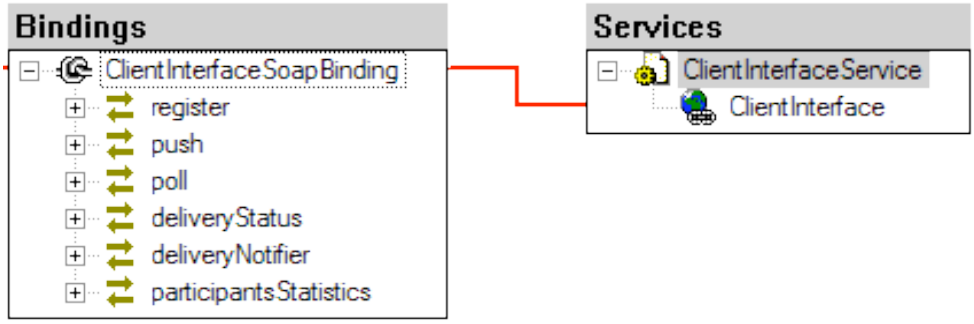


**Figure 1: WSDL bindings for client-to-peer communication**

As depicted in Fig. 1, we have defined a small set of SOAP queries that manage the interface between existing news agency services and network peers. The information flow from news agency to news agency can be controlled by corresponding SOAP queries targeted at the adjacent peer component.

The actual coupling of the single partners is provided by communication mechanisms between the individual peers. In our approach the systems only have contact to the local peer and exchange messages by requests and replies with each other. The request forwarding to other network partners is handled by peer-to-peer communication or loosely coupling of the symmetrically constructed peer components.

The communication protocol between the peers is very simple, messages are exchanged according to the sending and receiving semantics of the SOAP queries issued by the adjacent agency systems. The local peers handle information exchange, message queuing, data transfer security, and semantic relation management in between the transferred business news items (cf. section 4). Since data transfer has a strong commercial impact – who is sending business news onto what networks, how are the distribution costs calculated, documented and shared etc. – the communication channels are secured via a public key infrastructure (PKI).

For the current version a self-constructed peer-to-peer protocol based on java socket communication is used. The project consortium is testing in parallel a JXTA environment (cf. www.jxta.org) to compare with native P2P architectures in terms of performance and manageability.

# 4 Modern Content Connectivity Based on Semantic Coupling

The core research part aside of the modern service integration based on P2P architectures and the Web Service interconnection is the semantic relationship management between business news. Finding the most relevant business information that is semantically connected to a current news item or a specific user query significantly increases the quality of the service and holds the potential to boost service acceptance for even a higher commercial success. Within this section we define our approach on how to use semantic-based features of AI to retrieve most relevant related business news items and compose business news web documents to act as a hub to both its content and to other most relevant information items semantically close to its content.

## 4.1 News Retrieval and Semantic Content Relationship

The state of the art on Information Retrieval states [8] that the process of fulfilling the information needs of a searcher (user) is to answer the query of the user. In a distributed news domain this means that news relevant to the user query must be retrieved from the net of digital news partners. So, the query and the news has to be comparable during the retrieval process, i.e. formalisms to represent the content, metadata etc, have to be common or at least shareable. The integration of several agency services within a news distribution concept in our approach is based in the interrelation of three different processes:

(1) the update of news that stems from network partner to be further distributed or is uploaded to the semantic knowledge base,
(2) the retrieval of a set of news related to the incoming news, or
(3) the merge of the different set of news coming from different partners related to a concrete news item.

We call those three processes news update, news enrichment, and news merging respectively.
In order to express semantic relations the known Vector Space Model has been selected [14,17]. One strong feature of the Vector Space Model is to identify the proximity between a set of documents by weighting their most relevant (key)words and comparing them to each other. We use this feature to automatically identify the proximity of a business news item to the large archive of business news within the entire network on a per-language basis, so providers are implicitly offering the most relevant business news accompanied with the most relevant articles on the same topic and content automatically attached to his article.

For the practical use in our approach, a modification of the initial Vector Space Model has been implemented and evaluated. A specific module (*mAKEr*) is in charge to perform the distributed retrieval at the Nedine peers, even if the local existing service supports retrieval mechanisms. The content interrelation and the semantic aspects of the final results have to originate from the metadata of the news as well as from the process and calculations related to the words or sequences of the words included in the news and their relevance measure to the distributed process (update, enrich or merge). Different tests have been done in order to identify the best ways and performance of the distributed retrieval using worldwide used collections of texts [9]. The most promising of them are going to be applied to the set of news available in the evaluation of the model within participating news agencies (see section 5) in order to refine the prototype.

The *mAKEr* is the software that supports news update, enrichment and merging by maintaining a database of relevant information for news (or documents). In particular, the *mAKEr* performs three different operations: it extracts data relevant to a document upon uploading of the document, matches a document against the data collected from other documents to come up with a list of related documents for enrichment, which are ranked relative to their “similarity” with the enriched document, and merges several lists of related documents coming from different sources into a single ranked list. These operations correspond to automatic keyword extraction (AKE), document matching, and result set merging.

In order to perform document enrichment, some data has to be collected on the documents of a document repository. This data is collected by AKE [8] and stored in an internal database, which we call the Meta DB. The data consists of relevant keywords appearing in documents, together with their frequency of appearance, both in terms of number of documents in which it appears, and in terms of number of occurrences of the keyword in a document. Based on this data, a descriptor of the document is built and stored, which we call the document vector (since it is based on the Vector Space Model). A document vector has the document keywords as components, and a value for each component (the weight) which is calculated from the keyword frequencies. The Meta DB thus forms a “vector space” describing the documents from the document repository (provided that they have been uploaded).

Given a particular document, the process of extracting its relevant keywords is called keyword extraction. There are different possible definitions of what a keyword might be; for the time being, we can think of it as a word with proper meaning, that is, a word other than a preposition, an article, a conjunction, and the like. What is meant for a keyword to be relevant admits also several possible definitions. Automatic keyword extraction on a document occurs at the moment of uploading the document. The data is extracted from the document text and stored in the Meta DB. The data consists of the frequency of occurrence of each keyword in the document (the term frequency). Also, the number of documents in which each keyword occurs (the document frequency of the keyword) is updated for each keyword occurring in the document just uploaded. The weight of each keyword k in a document vector (for document d) is given by the following formula:

$$w_{kd} = tf_{kd} * idf_k$$
$$idf_k = \log ( N / df_k)$$

where $tf_{kd}$ is the term frequency of the keyword k in the document d and $idf_k$ is the inverse document frequency for the keyword k. $df_k$ is the document frequency for the keyword k i.e. how often k occurs in the entire document repository and N the number of documents in the document repository.

Additionally, weights can be normalized by dividing them by some factor (i.e. the length of the vector). The purpose of this is either to standardize the values to a given range, or to take into account other factors that might affect the effectiveness of document extraction.

The matching operation is the proper enrichment of a document. The document to be enriched is described by a vector as explained before and then matched against all other document vectors in the collection in the same language. Such a matching obtains a rank, which measures the degree of matching between the two documents. Documents with a rank above a given threshold are returned as the result set that enriches the given document. The result set is ordered by rank, which gives a measure of the relevancy of each document to the initial one.

The rank is a numeric value for the “similarity” between the two documents (i.e. news items) matched. The usual similarity measure used in document vector spaces is the inner product of the two normalized document vectors. The inner product is the sum of the products of the weights of each vector which correspond to the same keyword. This measure is also called cosine similarity.

In [3] a number of solutions for the merging problem in distributed information retrieval are proposed. All are based on formulae to recalculate the ranks of the documents in the ranked lists. Collection descriptors are assumed to be available. Having collection descriptors available, collections are ranked w.r.t. the query (the document to enrich, in our case). The ranks of the different collections will then be used in the merging. The main difficulty in applying such solutions to our case is that our collection descriptors cannot, in principle, be used to calculate ranks for collections.

The merging operation takes several ranked result sets and merges them into a single result set, where ranks correspond to a unique ranking for all documents. Since ranks measure the relevancy of documents to the document they enrich, the merging has to provide for the consistency of this relevancy measure in the final merged ranked list. In the business news area we assume that we have very similar vector spaces. That is, the vectors that describe documents in different collections which are similar (i.e., potentially related to a third document or between each other) are also very similar. In this case, the ranks obtained in the different collections are comparable. Thus, it is sensible to consider equal ranks as identical measures for relevancy. Then, result sets can be merged by simply putting ranked documents together (and sorting them accordingly to the document ranks).

However, if the above assumption does not hold true, then ranks have to be somehow recomputed when result sets are merged. In the following we review several possible solutions to the problem. We will assume that a collection is represented by the set of all pairs of a keyword occurring in a document of the collection and its document frequency. We call this set the collection descriptor. The set of all pairs of a keyword occurring in a document and its term frequency (or, the document vector) is called the document descriptor.

One possibility for merging result sets is to relate the documents they contain to a common collection of reference. The process can then act as if the documents were being uploaded into that collection, and their vectors in the vector space of the collection of reference are computed. We call this process re-indexing. Recalculating ranks: a number of solutions for the merging problem in distributed information retrieval were proposed. The main difficulty in applying such solutions to our case is that our collection descriptors cannot, in

principle, be used to calculate ranks for collections. One possibility, with the VSM, is to consider each of the distributed collections as a document in the (virtual) document collection of all the distributed collections. Some (sensible) approaches in this line have been considered for the prototype implementation (Nedine peers).

# 5 Lessons Learned: Integrating Semantic and Communication Technologies

The described approach has recently been implemented within the EU-funded project Nedine. Nedine is a news distribution network, initiated by leading news agencies in Austria, Germany, Switzerland, Slovakia and the Czech Republic and aims towards the integration of their news management and distribution networks in order to build a pan-European service for their subscribers and customers. The project consortium additionally includes European Universities (Technical University of Vienna and Universidad Politecnico de Madrid) and two major PR agencies in Germany and Switzerland.

## 5.1 Project Experiences

Quantitative background: The business model includes the creation of high quality information by editors as well as the distribution of PR news items for paying customers. In total, the Nedine project consortium has access to more than 200.000 subscribers with individually managed profiles and offers more than 100 business news daily backed by a news database of 120.000+ documents. Pre-project news distribution included approx. 580.000 emails per day and created 12,4 million page impressions per month on the participating news agency's web sites. The loosely coupled service integration by connecting the agency systems via peer components does scale very well and the project consortium estimates an increase in distribution traffic brought by the widened business opportunities for news distributors by 35% until 2007.

*Technical integration:* The existing services, i.e. the web presentations of the business news management and publishing as well as the electronic distribution management (mainly per email) has been standardized according to the requirements of the project consortium to participate in the Nedine network. Basically, identical products (news distribution, profile management) were identified and streamlined with as little technical implications as possible in order to reduce entry barriers for future network partners.

Technically, the introduction of a loosely coupled network between existing news agency services has proven feasible and was implemented in a prototype in Summer 2005 (see www.presstext.at resp. www.nedine.org). A completely new feature for the Nedine business news network are the semantic relations between news articles currently in focus and all relevant articles identified by the use of the similarity mechanisms described above. It is especially the combination of the modern connectivity of national news agencies to an international network with the semantic relationship management of interchanged and distributed news articles that really enhance the web functionality and bring practical utility.

*Content Relationship Management and Semantics:* Recently, some news agency partners presented hyperlinks from a document in focus to some semantically related documents. Those relations were either created automatically by relating the content topics that were assigned by the authors (category resp. topic neighbourhood) or were added manually by the editors, based on a preceding fulltext search (manual relationship management). The former approach has poor quality, since news items under the same major topic e.g. research need not necessarily be closely related with each other. The latter approach is very costly, since personal resources for the fulltext retrieval are necessary. Employing the semantic-based approach discussed in section 4, readers are always best informed with a service quality far beyond what could be reached by a single agency itself. The semantic relations created by a specific variant of the Vector Space Model (see section 5.1) added a new level of Web functionality to the integrated news presentation and distribution network Nedine: the ability to offer high quality interlinking of semantically related business news items.

*Prototype evaluation:* Early prototypes of employing the semantic-based Vector Space Model have presented highly acceptance results. An internal questionnaire amongst the news agency journalists has shown, that 68% of the experts are satisfied with the automatically created relations and 15% claim to be highly satisfied (total N=247 journalists, cf. Fig. 2). Semantic relations were created between totally 84.421 business news articles, the 5 most significant ones according to the similarity calculations within the Vector Space Model have been attached automatically to each article.

As discussed in section 4.1 specific characteristics of the application domain have to be considered to best control the similarity weighting. Currently the prototype is evaluated more thoroughly considering the following specifics of the business news area.

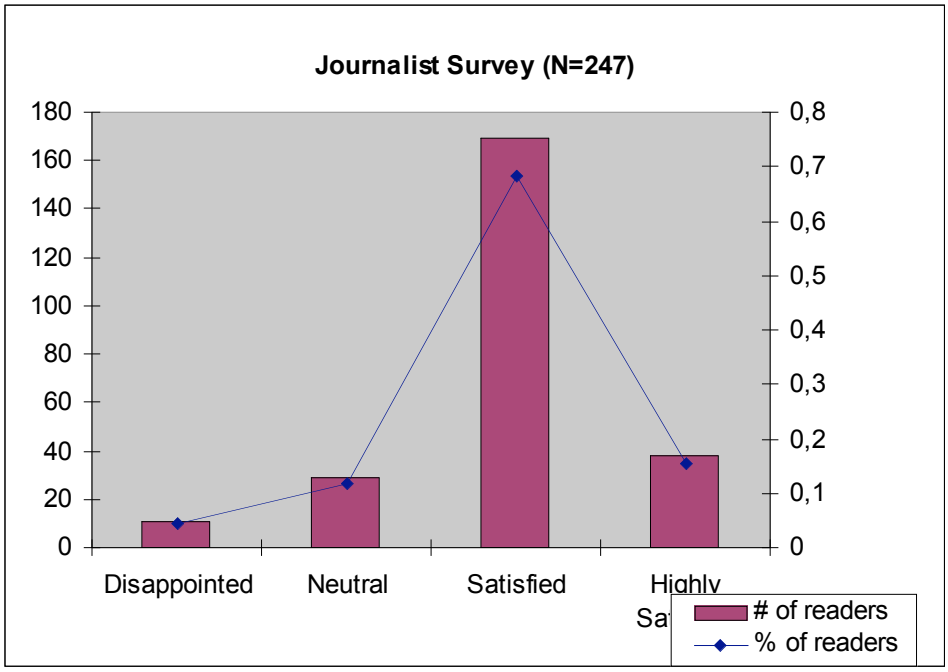


**Figure 2: Survey results on user satisfaction on quality of semantic relations between business news articles**

## 5.2 Semantic Relationship Applicability in the Business News Area

To customize and optimize the use of the Vector Space Model for Nedine and increase the quality of the news contents for both the readers and the distributors, two properties unique to the business news distribution domain have to be considered: The limited size and the typical structure of business news items:

(1) *Specific size of business news information*
In order to highly qualify the keywords of a news article and to set correct and significant values for the weights, the examined text should have at least a sufficient length. News items rarely hold more than 300 words, which is a rather small number for significant weighting. The adaptation we followed is to require a minimum appearance count for words to be registered as significant.

(2) *Specific structuring of business news*
The specific structure of news items add more parameters of significance to the quality of the Vector Space Model retrieval results. News items, although short in general, are structured in several sections such as title, subtitle, abstract, content, signature, attachments. We have extended the Vector Space Model by adapted weighting of keywords according to these specific structuring properties of news articles within Nedine.

For each structural element of the news item, individual weighting is calculated and similarity measures are created based on a vector of result values. This vector combined with an adaptable structure vector, holding weights as relevance indicators for specific sections of a news document, provides means of calculating the adapted significance resp. weighting for keywords in news documents. An example for factors within the structural vector is presented as:

$$Vstruct = \begin{pmatrix} 5 \\ 3 \\ 2 \\ 1 \end{pmatrix} \begin{matrix} \ldots \text{NewsML section Headline} \\ \ldots \text{Subline} \\ \ldots \text{Abstract} \\ \ldots \text{Content} \end{matrix}$$

which implies, that a keyword identified in the Headline section of NewsML is assigned five times the weight as if it was identified in the regular content section.

The prototype has been implemented partially at public available and private sections of the domain www.nedine.org to provide also a significant practical contribution to the area of Semantic Web. A commercial distribution of the designed platform is scheduled for Spring 2006.

# 6 Conclusion

The main goal in our research has been the creation of a business-oriented intelligent news publishing and distribution network, engaging modern Internet technologies such as Web Services, loosely coupling services via a simple P2P network and the employment of semantic-based methodologies to relate business news items to enrich the quality of the provided service. Beside the technological challenges of using state-of-the-art Internet technologies the approach is highly driven by industrial and commercial ideas. The research experiences from project partners have supported the creation of a well-defined service architecture, a communication and interaction model and specific network features that lead to a service result that is more than just the connection of existing business news distribution services.
XML expertise and system architecture experts for high performance distributed systems cooperate in this project with researchers in the field of semantic-based retrieval to provide a platform for highly qualitative multinational business information exchange. We have identified an adapted version of the Vector Space Model to semantically relate business news items and compose result documents containing not only the retrieved contents but also hyperlinks to the most relevant semantically related documents within an entire business news network, consisting of multiple participating national news agency repositories. The loosely coupling of strong existing services is introducing a new quality of business news presentation and distribution in Europe.
Technically, the network integration has been defined and its functionality and availability is proven in running prototypes, demonstrating the business opportunities at several running sites yet. Semantic relations have proven high quality in the evaluation of an editors questionnaire and provide a unique extension as a USP for future business implementation. The prototype has proven how semantics technologies like similarity calculations based on adapted Vector Space Models can provide new levels of Web functionality and enrich the service quality for business news readers as well as the business opportunities for news agencies in a practical environment.
Currently the prototype is implemented within the research work on Nedine, an EU-funded project with three European news agencies as initial network partners, currently covering seven countries in the Central and East European area. On successful further evaluation and the integration of additional network partners, the concept shall prove its quality and scalability within a real business environment, starting its productive phase in 2006.

# Acknowledgements

The current research work has been conducted in cooperation with experts of the distributed systems area at Vienna University of Technology, namely Prof. Shahram Dustdar and Christian Platzer, and the University of Zurich, Prof. Harald Gall. Semantic methodologies and Vector Space Model management has been evaluated and extended by the Universidad Politecnico de Madrid, Prof. Ana Maria Garcia Serrano and Prof. Francisco Bueno. This work was partially funded by the EU eContent project NEDINE (News Distribution Network, EDC-22225).

# References

[1] BIRMAN P. IEEE Internet Computing: Peer To Peer - The League of SuperNets. **In** *IEEE Distributed Systems* [online]. 2003, vol. 4, no. 10.

[2] BERNERS-Lee, T. *Semantic Web Road map.* http://www.w3.org/DesignIssues/Semantic.html, 1998

[3] CALLAN, J. Distributed Information Retrieval. *Advances in Information Retrieval* : Kluwer Academic Publishers, 2000, p. 127-150.

[4] CVITKOVICH, A. Applied Middleware: Object-Oriented Content Management Components with Mason. **In** *Proceedings of the IADIS Applied Computing Conference AC2005, Algarve, Portugal, February 21-24, 2005*.

[5] GUDGIN et.al. *SOAP version 1.2 (W3C)*. http://www.w3.org/TR/soap12-part1/, 2003

[6] HAAS, H. *World Wide Web Consortium - Web Services*. http://www.w3.org/2002/ws/, 2002

[7] IPTC, International Press and Telecommunication Council. http://www.iptc.org/

[8] Martinez, J. L.; GARCIA-SERRANO, A.; MARTINEZ, P.; VILLENA, J. “Automatic Keyword Extraction for News finder”, AMR 2003. *LNCS*, 2004, vol. 3094.

[9] MARTINEZ-FERNÁNDEZ, J. L.; GARCIA-SERRANO, A.; VILLENA, J.; MÉNDEZ, V. “Miracle approach to imageCLEF 2004: Merging textual and content-based image retrieval”. *LNCS*, 2005, vol. 3491, p. 699-708.

[10] NEDINE. – Intelligent News Distibution Network, EC eContent EDC 22225. http://www.nedine.org/, 2004

[11] NewsML,2003. IPTC-NewsML Standard. http://www.newsml.org

[12] PAEPEN, B. OmniPaper: Modern Approaches for an Intelligent European News Archive. **In** *Proceedings of the IADIS WWW/Internet2002 Conference (Lisbon, Portugal, November 13-15, 2002).*

[13] SALTON, G. *Introduction to Modern Information Retrieval, volume 1*. McGraw-Hill, Inc., 1993.

[14] SALTON, G.; BUCKLEY, C. Team weighting approaches in automatic text retireval, Readings in Information Retrieval. Morgan Kaufmann Publishers. 1998, p. 323-328.

[15] SCHRANZ, M.; PAEPEN, B. Architecture design and application for an Intelligent Distributed News Archive. **In** *Proceedings of the ICCC 8th International Conference on Electronic Publishing – ELPUB2004 (Brasilia, Brazil, June 23-26, 2004).*

[16] SCHRANZ, M. Employing Web Services and P2P Technology to integrate a Pan-European News Distribution Network. In *Proceedings of IEE ITA05, September 2005*, p. 461-468.

[17] WONG, W.Z.; WONG, P.; WONG, Patrick. Generalized vector space model in information retrieval. ACM, 1985.

[18] World Wide Web Consortium (W3C). Web Services Description Language (WSDL) 1.1, http://www.w3.org/TR/wsdl/., 2001

[19] XML iptc-subjectcode News Code. http://www.iptc.org/NewsCodes, visited Nov. 2005

[20] YU, S.; LIU, J.; LE, J. Decentralized Web Service Organization Combining Semantic Web and Peer to Peer Computing. *ECOWS 2004*. 2004, p. 116-127.