Personal Information Agent

Dominik Kuropka[1)], Thomas Serries[2)]

[1)] University of Muenster, dominik.kuropka@wi.uni-muenster.de
[2)] University of Muenster, thomas.serries@wi.uni-muenster.de

Abstract: *Information overflow is one of the greatest challenges for information focused professions today. This paper presents the Personal Information Agent, an agent based information filtering prototype. The prototype has been build to prove the concept of our agent and neuronal network based information filtering system which is presented in this paper. Further experiences made during implementation and ideas for future work are discussed.*

Keywords: *Agents, Information Filtering, Neuronal Network*

1. Introduction

Beginning with the use of computers more and more structured information were stored digitally. The more desktop computers (PCs) were used in offices the more they were used to create unstructured and semi-structured information like reports, letters, etc. More and more unstructured documents were and are generated using computers today. Development of the Internet supports this trend. Publishing documents is easier than ever before. The number of available documents/information is growing faster than ever.

By now information has become one of the most important resources for business and research. But the fact that more information is available for everybody than ever before, doesn't lead to better decisions. Restricted capacity of humans in information processing forces to reduce the amount of presented information. So today one of the greatest challenges is the efficient selection of relevant information: *information filtering* (IF).

In contradiction to being informed by information filtering systems users of *information retrieval* systems are searching for information actively. The user has an idea of the information she needs. She works with the information retrieval system and starts searching for information (ad-hoc query). Users of information filtering systems are informed about relevant information autonomously. So the main task of the system is to find out which of the collected information is good enough to support the work of the user without overloading her with useless information. [1]

2. Requirements

Information filtering systems have to meet at least the following four requirements to be of use in a wide range of use cases:

- The system should be able to observe several, different information sources on its own. To achieve this, it has to cope with different data and document formats. If information sources do not support push-mechanisms to trigger its listeners (e.g. web-sites), the information filtering system has to scan the information sources for changes regularly.
- Messages presumably important for the user should be presented at a glance.
- Information filtering systems will never reach 100% correctness in evaluating messages. So unimportant evaluated messages should be accessible, too.
- For being accepted by the user an easy-to-use user interface has to be provided by the system. A user should understand the main functionality of the user interface intuitively and unexperienced users should be able to customize their filtering profiles. Collecting information about user profiles should require as few interaction with the system as possible.

3. State of the art

Information filtering has two independent dimensions shown in the morphological framework in figure 1: the *classification approach* and the *classification method* used to implement the classification approach.

peculiarity	peculiarity value		
classification approach	importance based		keyword based
classification method	with pre−classification profile		without pre−classification profile
	explicit profile	implicit profile	

Figure 1: Morpholigical framework for IF

Referring to the first dimension, information filtering always requires some kind of classification of messages. In general two different classification approaches exist. The *keyword based approach* arranges messages depending on their content into one or more categories which are represented by keywords. Usually categories are structured within a hierarchy or network. The user has to select those categories she is interested in. *Importance based classification* assigns a degree of importance to the combination of each message and each user. This is an one dimensional numeric value derived from the content.

The second dimension describes two main types of classification methods. Information filtering systems *without pre-classification profile* in general (also called collaborative information filtering systems) sort information by date or by user activity or user voting. Alternatively information can be evaluated as important or appendant to a category if a quorum of other users marked it accordingly. Public examples for a systems implementing the method of importance based classification without pre-classification profile are the Linux Community website[1] or the NewsSIEVE[2] tool for Usenet newsgroups. Generally systems implementing classification without pre-classification profiles loose time between the appearance of a new message and the first time this message is classified by users. This approach is not user specific at all which makes it usable only, if all users have similar information demands.

In filtering systems profiles are used to represent users' informations demands. By asking the user about her information demand an *explicit profile* can be created. This strategy is often implemented in combination with the keyword based approach. Commonly known systems using keyword based pre-classification profiles are e-mail clients using filtering rules for example. Other examples are category based newsletters like the BDW-Agent[3] or news-tickers like Slashdot[4]. Users select categories which represent their own information demand at best out of a given set. New messages are categorized by editors and automatically send to all users who subscribed the appropriate category. Generally users have to cope with differences in the understanding of categories between themselves and the editor. This grievance becomes more obvious the more categories the system supports. To avoid this a reduction of categories is thinkable but leads to a less user specific system reducing the quality of information filtering.

Importance classification using explicit profiles is possible in theory but hard to realize in practice. Users of such systems have to define a rule set which exactly describes when they realize messages as important or not. The explication of knowledge is one of the hardest challenges for knowledge based information systems. Only experienced users reach sufficient quality. RAMA[5] is an example for an information filtering system using pre-classification with explicite profiles.

Independent from the classification approach one of the biggest disadvantages of information filtering systems using pre-classification with explicit profiles is the need to create user profiles before the system can work. For unexperienced users the manual creation of a user profile is difficult and not related to her work directly. Psychologically this means a high obstacle in using the system.

Alternatively user profiles can be created implicitly. Observing the interaction with the system adaptive algorithms or statistical analysis can be used to create user profiles. Benefits of this approach are the avoidance of explication of knowledge and the continuous adaption to changing information demands. The PI-Agent prototype described in this paper uses this approach.

Recapitulating, it can be state that public available information filtering systems exist, but they are using either user unspecific classification methods without pre-classification profiles or explicit user profiles. Available information filtering systems are isolated solutions for only one information source. Users have to cope with different environments and different classification approaches if they want to access news from different information sources.

4. Our work

The PI-Agent prototype implements an information filtering system using importance based classification realized by implicit pre-classification profiles. Followed by a de-

[1] http://www.linux-community.de

[2] http://www.newssieve.com

[3] http://www.wissenschaft.de

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

[5] ftp://ftp.cs.pdx.edu/pub/faculty/jrb/rama

tailed look at the functionality of the agent, the used tools and algorithms the architecture of the system is illustrated first.

4.1. Agent based approach

WOOLDRIDGE and JENNINGS argue that rational agents should have the following properties: autonomy, proactiveness, reactivity and social ability [12]. These properties make agents powerful and enable them to meet the requirements of adaptive information filtering systems. Possible specifications of these properties for information filtering systems are:

- *autonomy*: Information filtering agents should work in background; independent from the user. They collect news-messages and decide about the presumed relevance for their users. To support autonomy WOOLDRIDGE proposes constructs like beliefs, desires and intentions. [13] Information filtering agents estimate what might be of interest for their users. They desire to make correct decisions about the relevance of messages[6] and to keep their users informed about important news. Finally agents have a notion how to achieve their desires, e.g. by collecting user-feedback or direct questions to the user.

- *proactiveness*: To achieve their desires agents have to execute some actions supporting their aims, e.g. scanning news-sources or contacting the user directly to inform her about very important news.[7]

- *reactivity*: Using techniques of artificial intelligence (AI), e.g. artificial neural networks or statistical data analysis, agents adapt their owners' information demands and believes by feedback.

- *social ability*: Information filtering agents have to communicate with different news-sources as well as with users. The user-interface should be easy to use and give a feeling of working in cooperation with an intelligent being.[8]

4.2. Personal Information Agent

The *Personal Information Agent (PI-Agent)* provides information filtering functionality in an agent oriented manner as described above. It's architecture is shown in figure 2.

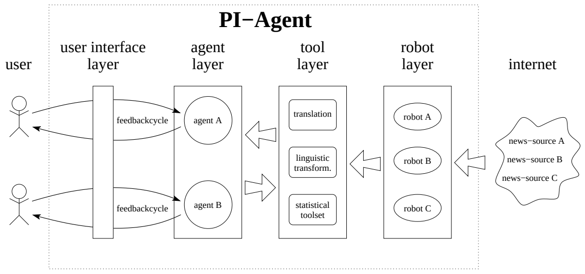

Figure 2: System architecture of the PI-Agent

News-sources are scanned for new messages by robots. To enable the PI-Agent to use as many news-sources as needed each robot is specialized for one (type of) source. Missing robots can be added during runtime. Currently the system scans the following web sites for new messages: internetnews.com, forbes.com, businessweek.com and Heise Online News[9].

To increase quality of presumption the second layer provides a set of linguistic and statistical tools as well as a translation service. The algorithms used in the tool layer are presented in section 4.3. in detail.

Users are represented by their agents in the third layer. The purpose of the agents is to estimate the user individual relevance of messages as good as possible. For this the agent passes all incoming messages to its artificial neuronal network. Currently the system is tested with a back propagation net simmilar to the net described by BOGER ET. AL. [2] To adapt its owner's information demand the agent collects message specific relevance evaluations given by its owner. Details about the neuronal network are discussed in section 4.4.

The user interface layer uses the estimated relevance evaluations of the messages to present them to the user in an appropriate way. To enable the agent to learn from user's evaluation of messages the interface has to provide feedback mechanisms. A HTML interface is realized but others are possible: e.g. integration into applications like mail clients and groupware systems as well as the usage of protocols like SMTP or SMS.

4.3. Language Processing Toolset

Categorizing of information represented by natural speech means recognizing of similarities between documents. So numerical representations of documents have to be generated to use similarity recognizing methods

[6] Agents desire to make the same decision about relevance of messages like their users.

[7] For example Short Message Service (SMS) or e-mails can be used.

[8] The user has to be aware of the fact that an intelligent being has to learn how to solve its tasks properly and that mistakes might happen.

[9] http://www.heise.de/newsticker

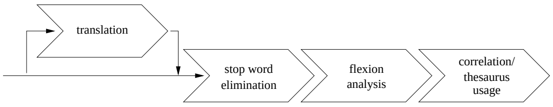

Figure 3: Steps of linguistic processing

known from statistical data analysis. Researches in computer linguistic developed toolsets which are used within the PI-Agent architecture. Figure 3 illustrates linguistic analysis of texts.

The first step of linguistic processing is to eliminate so called stop words form further processing. These words are used very often within all documents and the frequency of usage is nearly constant within different documents. Because of this they do not give any information about the content of a certain text. Examples for stop words for English are: “and”, “the”, “very”, “by”, or “which”.

Further reduction of data can be achieved by using the knowledge about language specific flexion of words. One word (interpreted as meaning) can be written in different sequences of letters: e. g. ‘misunderstand’ has the same meaning as ‘misunderstood’. Difference in time does not influence its meaning. Depending on the used algorithm reduction may lead to loss of information. So two different words may be represented by the same term (e.g. ‘mine’) or terms of different words are reduced to the same basic form. The problems resulting from this can only be solved by methods of semantic text analysis. A collection of different approaches can be found in [9]. None of them is implemented in the PI-Agent yet.

The categorization of documents requires the recognition of their meaning. Analyzing basic forms of the used words ensures to recognize different spellings of one word as the same meaning. But different words (basic forms) may have the same meaning or one word is often (only) used in conjunction with an other. To avoid negative effects of correlated words a statistical toolset is used. Alternatively the linguistic analysis can use thesauri to eliminate problems resulting from correlated words. Thesauri may enable further enhancements if additional information like generalization and specialization of words are also taken into account.

Because stop words and flexion analysis are determined by the language these tools have to be provided for every supported language. Availability and quality of these tools are language dependent, too. To avoid these dependencies texts written in a not supported language are translated into the target language. Depending on the quality of the translation tool this may lead to major problems in recognizing the correct meaning of texts. For commercial use the following aspects have to be taken into account for decision if this solution is suitable:

- *How is the translation tool working?* Translation tools doing a simple word-by-word translations may tend to false interpretations of some words. Because the PI-Agent does not perform linguistic but statistical analysis, resulting errors will be of minor importance. Wrong translations will be made the same way always. Translation tools based on linguistic analysis will have to be domain specific because they may translate words depending on the context into different words of the target language.

- *How many domains is PI-Agent used in?* Depending on the domain of a text one word of the source language may have different translations in the target language. If PI-Agent is used in heterogenous domains this means that translation tools have to be able to recognize the domain of texts and use the corresponding dictionary.

Because usually one message is evaluated by more than one agent it is useful to execute the translation and the stemming algorithms only once for reasons of performance. Focusing on the main research target the decision was made to implement the linguistic analysis for English only. Used algorithms and methods are taken from [5] (stop word list), [14] (stemming rules), and [6] (correlation analysis). Support for non-English documents is realized by the integration a freely available translation service (www.babelfish.com).

4.4. Agents and Neuronal Networks

User specific agents provide the core functionality of the PI-Agent system. They represent user profiles and evaluate messages. Adapting themselves to users’ information demands and using linguistic and statistical tools increases quality of relevance presumption.

The evaluation mechanism itself is implemented as an artificial neural network [7]. The processing elements (neurons) of agent’s neuronal network are perceptrons [8] which are organized in input, output and hidden layers. Similar to the approach described in [2] each neuron of the input layer represents one keyword. For each keyword being part of a message the corresponding neuron is set to the value 1, all other neurons are set to the value 0 (figure 4). After processing the input the neuronal network computes the result in the output layer which contains only one neuron. The result range [0;1] of the neuron is translated into human readable categories from one

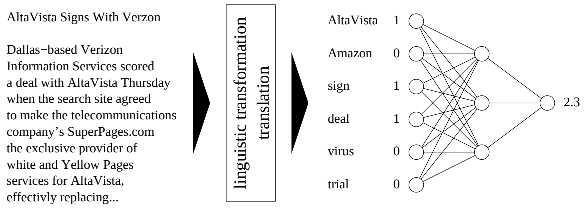
AltaVista Signs With Verzon
Dallas−based Verizon
Information Services scored
a deal with AltaVista Thursday
when the search site agreed
to make the telecommunications
company's SuperPages.com
the exclusive provider of
white and Yellow Pages
services for AltaVista,
effectivly replacing...
linguistic transformation
translation
AltaVista 1
Amazon 0
sign 1
deal 1
virus 0
trial 0
2.3

Figure 4: Evaluation of information

to six which represents the predicted, user individual importance of a message.

Every message can be evaluated by the user to give feedback to her agent. The agent collects this information and starts the adaption processes regularly. Currently the last 150 user evaluations are used to recreate and train the neuronal network by the back propagation algorithm as described in [7].

For implementing artifical neural networks the programming paradigm matching the requirements, structure and dynamics at best should be chosen. Neuronal networks consist of a set of data representing the structure of the network and of a set of algorithms working on the data and representing its behavior. Objects – which are the key-components of the *object oriented approach* – have the same properties: They consist of data which is encapsulated by the object. Methods implement interfaces to object's data and define its behavior, i.e. all other functionality to process its data. [3] So, the object oriented approach – with its distinction between object's structure and behavior – fits good into the implementation requirements of neuronal networks.

Just representing the network by an object does not lead to the highest degree of abstraction. For information filtering the neuronal network alone is not sufficient. Linguistic and statistical algorithms are needed as well as access to the training data. The training data is needed for the adaption mechanism being executed autonomously and not necessarily synchronized with other system components. To abstract from this details and to allow reusability and flexibility a higher level of abstraction is needed: agents. Section 4.1. already showed that the agent concept provides all functionality needed to encapsulate the complexity of an information filtering neuronal network.

4.5. Implementation details

The PI-Agent prototype is fully implemented in *Java*. Persistency is ensured by *PostreSQL*[10] database management system with *JDBC* as communication interface between agent, user interface and database. The web-based user interface[11] runs on an *Apache*[12] web-server, dynamic parts of the web-site are implemented using *Java-Servlets*.

5. Use cases

By now news agencies provide information for their customers by pull mechanisms or by push mechanisms neglecting customer individual information demands. Messages are generated and classified according a given scheme. To overcome difficulties of the category based classification described in section 3. news agencies can use PI-Agent systems: Generated messages are no longer classified. They are published in a prohibited area, so only the customers (and their agents) can access. Agents are checking for new content regularly and inform their owner whenever relevant information is found. Customers pay per viewed message. Alternatively news agencies can provide messages as they do today. As additional feature they can use the PI-Agent technology to keep track of customers' information demands. Every time a message is evaluated as important for one customer by its agent the message is sent to the customer using a separate channel (e.g. SMS or Mail). If standardized interfaces for agent communication are provided by news agencies the first alternative enables customers to use one agent for several agencies. For the second alternative every agency has to host one agent for each customer.

The PI-Agent collects information about the information demands of its owner. So consultants can use PI-Agents to observe employees doing their daily work. Afterwards all possible news sources can be evaluated by the agents which may give advice for the best selection of information sources. The PI-Agent may recognize information sources with a high degree of overlapping.

Often Internet portals provide the option of informing their users (regularly) about new product offers by mail. To increase efficency of such advertise mailings user may be observed by their agents. Using the collected information about interests of their owners the agents may forward only interesting offers. Alternatively the agent may scan all Internet sites of the portal for potentially interesting pages. Because users interests may change she will be informed about information in not viewed parts of the portal automatically. Also users may be informed when-

[10] http://www.postgresql.org

[11] http://www.pi-agent.com

[12] http://www.apache.org

ever new Internet sites of potential interest are inserted into the content of the portal. While mail filtering requires hosting agents by the Internet portal, tracking of the content of the portal can (but does not have to) be realized by external agents.

Within enterprises the PI-Agent technology can be used to realize user-individual bulletin-boards. Exchanging knowledge by bulletin boards is a powerful way to share business knowledge within the whole enterprise. As stated in section 3. searching and classification of information is not easy. So such systems often lack of acceptance by the users. Supporting them in information retrieval can help to overcome these retentions.

Generally PI-Agent systems can be used to create individual information sources (using push mechanisms) out of a set of non-individual information sources (using pull mechanisms).

6. Related work

Several researches on information filtering systems have been made in the past. In this section two of the most interesting projects are compared to the PI-Agent system.

MINT is an abbreviation for 'Management Information from the Internet': It implements a prototype of an editorial workbench. [4] The key concept is the support of two different user groups within an enterprise. The first group consists of in-house information brokers who provide news to the other group – the information receivers, such as managers. The main tasks of the MINT system are collecting information from different web-sites, supporting information brokers in evaluation and categorization of messages and enabling adequate presentation of information to information receivers. The main difference to the PI-Agent is the use of human resources for evaluation of information. Potentially a better presumption quality is reachable but raises the total costs of the system. The engagement of information brokers in small or medium sized business is often more expensive then the achieved improvements reduce costs.

As described in section 4.2. the agent layer of the PI-Agent architecture contains a artifical neural network as described by BOGER ET. AL. [2] The authors employed their neural network for information filtering and term selection. About 1500 e-mails from 10 users where used to train the networks. With this data they reached prediction quality of 76–99%. SHAPIRA ET. AL. found out that "content-based filtering" and "sociological filtering" (and combinations of them) only reach 40–70% prediction precision. [10] Content-based filtering is based on correlation of two weighted vectors of terms (one for the user and one for the information). Sociological filtering defines user groups basing on evaluation profiles of the users. The relevance evaluation for one user is made from her own evaluation rules and the rules of the corresponding user group in parallel. Comparing quality of these approaches information filtering using neural networks shows significant advantages.

7. Discussion and future work

The aims of the PI-Agent project are to prove the suitability of the described architecture for information filtering. For evaluation of presumption quality data from four regularly users of the PI-Agent system were analyzed: The last 150 user-evaluated messages from each user were used to train one neural net for each user. For a test set consisting of 254 messages not used for training a presumption precision of 80% with a standard deviation of 11% was reached. A message pre-evaluated by the system was rated as correct if the difference between user evaluation and system evaluation was equal or less than one grade. Starting points to improve quality are:

- The representation of keywords in input layer of the neuronal networks should be changed from $\{0;1\}$ to $\{-1;1\}$. This makes neuronal networks more powerful. So, the following situation can be reflected by networks: 'A message is important if keyword x but not y is part of the message.'
- Other training algorithms could improve learning and enable re-usability of networks (maybe in combination with genetic algorithms).
- Thesauri transform used words into their meaning. This reduces correlation between choice of word and evaluation.
- Integration of pre-classification without profiles based on statistical analysis and using explicit profiles to allow users to describe what information they miss may reduce pre-evaluation errors.

In discussion with users of the system the user interface has been the main reason for critics. The system should not require feedback for doing its job. It should record users' behavior to collect information about their information demands. So the user will only have to give feedback in case pre-evaluation was wrong.

On the technical level the agent concept is adequate for abstraction from complexity of the information filtering problem. But for a seamless implementation powerful agent platforms are needed. A good platform should meet at least the following criteria:

- *Transparent persistency support*: The neuronal network represents agent's knowledge about user's information demand and is created by a long running process of communication with the user. Data stored within the network is necessary for the system and hard to restore if a system crashes. A transparent persistency layer ensures that agents can be recovered after crashes or after structural changes of the agent-software. It might be usefull to adapt persistency methods known for object oriented systems [11] into agent technology.

- *Scalability*: The resource usage is mainly depending on the number of agents running. Agents are independent from each other; but PI-Agent architecture provides layers that might be used by them in parallel. To avoid implementing same functionality several times the agent platform should be scalable at least for these layers.

8. Final remarks

As shown in this paper the PI-Agent is an adequate starting point for further research in field of information filtering. Insufficient presumption quality can be enhanced by justification of the training parameter of the neural network, implementation of other learning algorithms, addition of further linguistic tools or integration of other classification methods. Never the less the PI-Agent architecture enables an abstraction from implementation details of the evaluation component. The reached level of abstraction is sufficient to integrate the PI-Agent into usual agent platforms or systems.

The PI-Agent systems may even be integrated into other systems like enterprise information portals which do not manage qualitative information only. Here the PI-Agent may be able to support the user in navigating through the set of reference objects, performance figures and reports. Even seamless integration of qualitative and quantitative information might be possible.

References

[1] Belkin, N.; Croft, W.: *Information Filtering and Information Retrieval: Two Sides of the Same Coin?*. In Communications. of the ACM, 35 (12), pp. 25-33, 1992.

[2] Boger, Z.; Kuflik, T.; Shapira, B., Shoval, P.: *Information Filtering and Automatic Keyword Identification by Artifical Neural Networks*, In: H. R. Hansen, M. Bichler, H. Mahrer: Proceedings of the 8th European Conference on Information Systems (ECIS 2000), Volume 1, Vienna 2000, pp. 379–385.

[3] Booch, G.: *Object-Oriented Analysis and Design With Applications*. Addison-Wesley Pub Co, 1994.

[4] Meier, M.: *MINT - Management Information from the Internet*. http://www.wi1.uni-erlangen.de/projects/mint /mint.pdf; downloaded on July, 30th 2001.

[5] Drott, M. C.; *A Big Stop List*; http://drott.cis.drexel.edu /retrieval.html; downloaded on July, 30th 2001.

[6] Kurbel, K.; Szulim, D.; Teuteberg, F.; *Künstliche Neuronale Netze zum Filtern und Klassifizieren betrieblicher E-Commerce-Angebote im World Wide Web - eine vergleichende Untersuchung*; Wirtschaftsinformatik 42(2000) 3, S.222-232

[7] Principe, J. C.; Euliano N. R.; Lefebvre, W. C.: *Neural and Adaptive Systems: Fundamentals through Simulations*. New York, 2000.

[8] Rosenblatt, F.: *The perceptron: A probabilistic model for information storage and organization in the brain*. In: Psychological Review 65, 1958, pp. 386–408.

[9] Salton, G.; McGill, M. J.: *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[10] Shapira, B.; Shoval, P.; Hanani, U.: Experimentation with an Information Filtering System that Combines Cognitive and Sociological Filtering Integrated with User Stereotypes, *Decision Support Systems*, 1999.

[11] Weske, M.; Kuropka, D.: *Flexible Persistence Framework for Object-Oriented Middleware*. http://www.kuropka.net /Dateien/Flexible_Persistence_Framework.pdf, 2001; downloaded on July, 30th 2001.

[12] Wooldridge, M.; Jennings, N. R.: *Intelligent agents: Theory and practice*. The Knowledge Engineering Review, 10(2):115-152, 1995.

[13] Wooldridge, M.: *Reasoning about Rational Agents*. London, 2000.

[14] Zhao, J.; *CS6704 Domain Engineering and Systematic Reuse: Class Project – Conflation Domain Engineering*; http://csgrad.cs.vt.edu/˜jxzhao/6704/project; downloaded on Juli, 30th 2001.