REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
05-06-2003	Final Report	7 November 2002 - 07-May-03

4. TITLE AND SUBTITLE

Interface For Fusing Human And Robotic Intelligence Using Scale-Free Small World Structures

5a. CONTRACT NUMBER
FA8655-03-1-3084

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Professor Andras Lorincz

5d. PROJECT NUMBER

5d. TASK NUMBER

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Neumann János Számítógép-tudományi Társaság
and Eotvos Lorand University
Pazmany Peter setany 1/C
Budapest H-1117
Hungary

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0014

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
SPC 02-4084

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Report is 15 pages. Electronic file also contains software demonstration of information fusion techniques.

14. ABSTRACT

This report results from a contract tasking Neumann János Számítógép-tudományi Társaság and Eotvos Lorand University as follows: The contractor will perform exploratory research to investigate scale-free small world network (SSW) technology. The techniques developed could be extended to modular Bayesian inferencing. As described in the technical proposal, the utility (effectiveness) of the approach will be investigated using two representative problems:
I: Extraction and Representation of Semantic Information based on Word Associations
II: Semantic relations amongst medical documents

15. SUBJECT TERMS

EOARD, Information Fusion

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18, NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Paul.Losiewicz, Ph. D.
UNCLAS	UNCLAS	UNCLAS	UL	15 plus SW demo	19b. TELEPHONE NUMBER *(Include area code)* +44 20 7514 4474

INTERFACE FOR FUSING HUMAN AND ROBOTIC INTELLIGENCE USING SCALE-FREE SMALL WORLD STRUCTURES

FINAL REPORT
CONTRACT: FA8655-03-1-3084

Contact: Mr. István Alföldi, NJSzT
PI: dr. habil. András Lőrincz

Budapest

May, 2003

Table of contents

Declarations

Acknowledgment of support and disclaimer have been added to all publications related to Contract No. FA8655-03-1-3084 in accordance with 252.235-7010 as described below:

252.235-7010 ACKNOWLEDGMENT OF SUPPORT AND DISCLAIMER (MAY 1995)

(a) The Contractor shall include an acknowledgment of the Government's support in the publication of any material based on or developed under this contract, stated in the following terms: This material is based upon work supported by the European Office of Aerospace Research and Development, Air Force Office of Scientific Research, Air Force Research Laboratory, under Contract No. F61775-00-WE065

(b) All material, except scientific articles or papers published in scientific journals, must, in addition to any notices or disclaimers by the Contractor, also contain the following disclaimer: Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the European Office of Aerospace Research and Development, Air Force Office of Scientific Research, Air Force Research Laboratory.

DATE: May 8, 2003
Name and Title of Authorized Official: _____"
Andras Lorincz

THE FINAL REPORT DELIVERED UNDER THIS CONTRACT SHALL INCLUDE OR ADDRESS THE FOLLOWING TWO ITEMS:

(1) In accordance with Defense Federal Acquisition Regulation 252.227-7036, Declaration of Technical Data Conformity (Jan 1997), *All technical data delivered under this contract shall be accompanied by the following written declaration:*

"The Contractor, NJSzT, hereby delcares that, to the best of its knowledge and belief, the technical data delivered herewith under Contract No. FA8655-03-1-3084 is complete, accurate, and complies with all requirements of the contract.

DATE: May 8, 2003
Name and Title of Authorized Official: _____"
Andras Lorincz

(End of Clause)

(2) In accordance with the requirements in Federal Aquisition Regulation 52.227-13, Patent Rights— Acquisition by the *U.S.* Government (Jun 1989), *CONTRACTOR WILL INCLUDE IN THE FINAL REPORT ONE OF THE FOLLOWING STATEMENTS:*

(B) "I certify that there were no subject inventions to declare as defined in FAR 52.227-13, during the performance of this contract."

DATE: May 8, 2003
Name and Title of Authorized Official: _____"
Andras Lorincz

(End of Clause)

1. Description of Tasks

The tasks of the contract were as follows:

Problem I: Extraction and Representation of Semantic Information based on Word Associations

Recent results in the literature indicate that synonyms and associative maps amongst words form scale-free small world (SSW) structure. Traditional graph representations seem to be inefficient to visualize scale free structures. Efficient representation is to be developed by using coordination numbers (the number of incoming and/or outgoing connections) at different scales. The program will be written in C++, visualization map editor and map viewer software will be provided in JAVA.

Problem II: Semantic relations amongst medical documents

Clustering of different medical sites, like American Heart Organization will be analyzed using SSW similarity measures. These studies will focus on the different parameterization and the optimal representation for humans. It is assumed – according to psychological results – that the optimal representation has 7 ± 2 nodes on each "screen" of the map. Additional maps will be developed by traditional factorization methods. Medical experts will judge – without knowing the algorithm – which type of information is better. These studies will be conducted by asking the expert to use the maps and measuring the time they need to find the information.

Problem I is a technology development. It was fully accomplished. The goal of Problem II was to measure the usefulness of the technology. During the course of the study, it was found that *yes*, the navigation on the maps is much easier. These maps can be studied at the following site: http://www.coraler.com/ (username: coraler, password: qwerty)
There are different maps to demonstrate the principle. For example, one map is about the World Intellectual Property Organization (WIPO), another one is about American Heart Association (AHA).

During the course of this SSW study, we found that the synonyms and the hypernyms barely help in grouping/clustering the information. Given that this direction is a crucial point in our future plans, I have added extra manpower to this work and we have executed additional research – not supported by the USAF project. This modification was necessary to have confidence in our IRI project. This additional project is referred as a pre-study below. In this pre-study automated keyword and key-phrase extraction methods were implemented. That allowed us to better characterize the maps. Now, there are two types maps that can be viewed on the site www.coraler.com: Some maps are without keywords, others have keywords. The AHA site *has* keywords and key-phrases. The WIPO site does not. The WIPO map, which has keywords and key-phrases was developed using the knowledge gained on the AHA site. One can view this as a description of the WIPO site using the knowledge gained on the AHA site.

Problem II was modified to a pre-study.

The question of this pre-study was,

- which clustering
- which key-phrase extraction

is the best on the different databases to select pages of interest. Is there a winning algorithm, or combination of algorithms are needed for fusing human and robotic intelligence.

Two basic methods were tried:

- FAQ and Answer

  Words of frequently asked questions and their hypernyms were chosen by known techniques getting rid of frequent words such as `the' `a', etc. Searches for pages with similar words were executed. It was studied, how well the words of the found pages matched the words of the answers belonging to the respective FAQs.

  - evaluations were executed with and without hypernyms
  - different databases were studied

  This was a brute-force study to establish 'expert – non-expert dictionary' (EnE dictionary). The overall result is that hypernyms by themselves might help. In most cases, however, our methodology can be questioned because both FAQs and answers were written in similar style. The set of hypernyms provided by Word Association is for general use and does not have the appropriate synonyms on novel technical -- medical fields. Different approach is needed to develop 'expert – non-expert dictionary' dictionaries. There are novel algorithms making use of Internet documents, which can be used to replace hand made collections of hypernyms and which are up-to-date. Such methods need to be investigated.

- Clustering and Keywords

  Different clustering and different key phrase extraction methods were used to find words in each database, which are representative for the clusters. In other words, different clustering methods were evaluated on the basis if the clustering was – up to some extent – topic dependent. It was found that by all means, Page Rank is not appropriate for such clustering; directed clustering methods produced much better results.

First, a short summary is given. This summary is followed by detailed description of the motivations. Overview of the results is provided in the Section 4. Details of the algorithms, software codes, results are provided in the Appendices:

Content of Appendices:

(1) **Appendix 1**: **Technical notes**
   (A) Literature in html form
   (B) Manuscript of a parallel project on link-highlighting, a direct human-computer interaction.
   (C) Manuscript of a parallel project on breaking news detection. It supports our claim that reinforcement based human-computer interaction and scale-free representations *together* can be very efficient. Moreover, it demonstrates that work-sharing amongst reinforcement learning agents can be achieved without direct communications, a major advantage.

(2) **Appendix 2: Databases, extracted keywords and key-phrases**
   (A) Internet databases used in our analysis, including the html pages of the
      (a) American Heart Association (called "AHA": http://www.americanheart.org)
      (b) Center for devices and radiological health (called "General": http://www.fda.gov/cdrh/index.html)

(c) Collaborative hypertext of Radiology (called "Radiology": http://chorus.rad.mcw.edu)

(B) A note, which describes the automated clustering algorithms.

(C) Clustering using connectivity properties
- (a) Outgoing link structure ("outdeg"),
- (b) Outgoing link structure with removal of selected nodes ("greedy outdeg")
- (c) Incoming link structure ("indeg")
- (d) 'Classical' PageRank ("Page rank")

(D) Extraction of keywords and key-phrases
- (a) keywords and key-phrases found on the html pages
- (b) question evaluation: the FAQ-answer method
- (c) key-phrases extracted by AI methods such as
  - (i) the KEA algorithm and
  - (ii) the mutual information method

(E) Pseudo code of each algorithm applied

(3) **Appendix 3: Detailed results on expert-non-expert dictionary**. This point was to replace the original suggestion that medical experts will judge – without knowing the algorithm – which type of information is better. Such studies seemed too much uncertain and were replaced. The new test is as follows:
- (a) Take non-experts as Frequently Asked Questions
- (b) Compare found pages using (A) words of FAQs and (B) words of FAQs extended by Word Associations
- (c) Compare the content of found pages with the content of the answers to FAQs

(4) **Appendix 4: Detailed results on keyword and key-phrase extraction.** Investigation of efficiency of keyword extraction methods and clustering methods: Which keyword extraction method works best with which clustering method to localize / select documents of interest in different clusters with high probability

(5) **Appendix 5: Software codes**

(A) Software of hierarchical clustering
- (a) Exe file of the C++ software, which performs the hierarchical clustering
- (b) Source file of the C++ software
- (c) Html documentation

(B) JAVA source code of the MapEditor, MapViewer and the Portal. This is the core software for human computer interaction.
- (a) An exe file, which installs the MapEditor
- (b) A set of maps made from the Word Association database. These maps can be *navigated and edited* by the MapEditor. Use simple drag-and-drop feature to edit the maps.

(6) **Appendix 6: KEA key-phrase extractor**

(A) Paper describing KEA

(B) Software, which is free for *academic use*
- (a) source code
- (b) executable
- (c) terms of licensing

The full report is about 300 MB and it is on the CD, attached to this rtf file. Some materials, but the software can be downloaded from http://people.inf.elte.hu/lorincz/Files/USAF/Final_Report

2 Summary

Methods

All the methods are detailed in Appendix 2. Algorithmic details and pseudo codes are provided in that part of this Final Report.

Main algorithmic components

Top-down clustering method:
The SSW structure is used to provide the hierarchical graph representation of the original graph with the smallest possible diameter and with the best match to the original network. Methods include o PageRank™ (method used by Google) clustering o Clustering based on incoming links o Clustering based on outgoing links o A 'greedy' version of clustering using outgoing links *It was found that all clustering methods have their merits but, interestingly, PageRank™ was the worst amongst the different clustering algorithms.*

FAQ and answer based evaluation using hypernyms
The study compares if words of FAQs or if words of FAQs extended by their hypernyms are better for finding html pages which match the words provided in the answers to those FAQs. *It was found that this brute force EnE dictionary has minor advantages if applied alone*

Keyword and key-phrase extraction method
Different keyword and key-phrase extraction methods were tried: o The KEA algorithm o Words of the 'keyphrase' metatag o KEA extension of key-phrases

Medical sites studied

Clustering, keyword and key-phrase extractions were executed on the following sites
American Heart Association (AHA): http://www.americanheart.org Center for devices and radiological health (General): http://www.fda.gov/cdrh/index.html Collaborative Hypertext of Radiology (Radiology): http://chorus.rad.mcw.edu *Details can be found in Appendix 2*

The main conclusion of this pre-study:

All of these state-of-the-are algorithms have their own promises, none of them is satisfactory by itself, and that user specific fusing requires goal-oriented combination of these (and similar) methods. That is, alike to the link-highlighting solution (Appendix 1.C) reinforcement learning has its promises here.

3 Motivation of the work

There is a need for finding information efficiently. There are several problems here:

1. The information is growing at an enormous rate
2. Today, not the information but the refreshment of the information is the key
3. Information is many faceted. Sometimes obvious things can not be understood without an appropriate dictionary
4. Dictionary could mean glossary or thesauri. These are complementary tools,
    a. the basic element being the glossary, i.e., the explanation of the terms
    b. the advanced element being the thesaurus

The desired *object* is the *conceptual graphs*, whose building blocks are *concepts* and *conceptual relations*.

No wonder that different efforts have been initiated, including

1. semantic networks and
2. topic maps.

Topic maps, for example, put the emphasis on *topics, associations, and occurrences*. XML forms of topic maps, HyTM and XTM, have been developed recently. However, although these technology components could be of importance, there is doubt that they will be able to conquer the world. In fact, the impression is that such technology components develop at a rate that nobody can become familiar with any of them before it becomes obsolete.

Let us see what they say about topic maps[1]:

„… The genesis of topic maps is to be found back in the early 1990's when what later became known as the Davenport Group was discussing ways of enabling the interchange of computer documentation. The group went on to develop DocBook ([DocBook 1999]), one of the most widely used DTDs for authoring SGML and XML documents. One of the problems the Davenport Group faced was that of merging the indexes of different sets of documentation, and the insight they arrived at was that:

> Indexes, if they have any self-consistency at all, conform to models of the structure of the knowledge available in the materials that they index. But the models are implicit, and they are nowhere to be found! If such models could be captured formally, then they could guide and greatly facilitate the process of merging modeled indexes together."

It is our firm belief that this note is very true. Yes, topic map is a useful tool for a well established subject: *when we know* what we are after and when we can find that material. In this case, a good topic map allows us to glance through the material quickly and efficiently.

By the same token, however, it is our firm belief that

> there is no reason to expect that constructs like topic maps for *novel* and for *not yet found information* could be of use by any means.

This pre-study is intended to be a forerunner of our three year project to be financed by USAF.

[1] S. Pepper. The TAO of Topic Maps: Finding the Way in the Age of Infoglut, http://www.ontopia.net/

The goals of that three year projects are:

- to make robots to search for information,
- allow for interaction between robotic and human intelligence in order to
    - improve searching and filtering capabilities of the robots
    - make robotic information amenable for human intelligence
    - adjust this collaborative work to the topic at hand, to the information available and to the intelligence and knowledge of the human

In other words, the goal of the research is to develop topic maps

- given the actual and not yet known *topics*,
- making use of the *associations* of the human and discovering the *associations* in the materials found by the robot
- fusing these *associations* to meet each other, and
- finding the novel *occurrences* given the fused knowledge.

This process could be seen as a topic map jointly developed by robots and human(s). The goal of this interplay is to make novel information amenable for human intelligence.

The underlying assumptions are as follows. It is assumed that

1. The connectivity structure (associations, links on the Internet) is not random, it represents some semantic overlap.
2. The connectivity structure is *rational*: it is the results of a selective process.
3. Selective processes give rise to scale-free small world structures.

Assumptions 1 and 2 are reasonable. Assumption 3 has many supporting evidences on all type of networks found in biology, neurobiology, social relations, relations in thesauri, etc. The abundance of the examples and the rare occurrence of counterexamples make Assumption 3 attractive.

The last formulation of our goals is as follows:

Given the connectivity structure our goal is

- to cluster the structure and
- to find indices describing *how* or *why* are the clusters related.

In our pre-study, we aim to find *indexing words*, which are representative to the clusters.

4 Results: An Overview

Results of the 'FAQ – answer' method

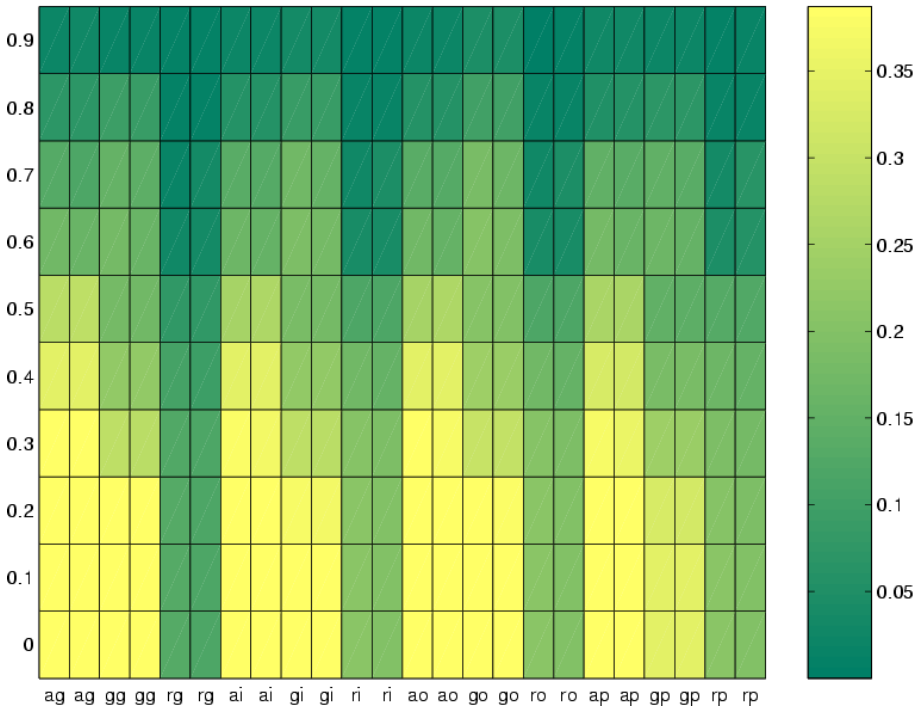

Figure 1. Evaluations of the FAQ—answer method
Database: All three databases. First letters of the horizontal axis: a=AHA, g=General, r=Radiology
Results are averaged over clusters and FAQs
Vertical axis: threshold values
Clustering method: All clustering methods: Second letters of the horizontal axis: g=greedy_outdeg, i=indeg, o=outdeg, p=PageRank
Column belonging to the first occurrence of the FAQ number: questions *without* hypernym augmenting
Column belonging to the second occurrence of the FAQ number: questions *with* hypernym augmenting
Colors represent the squared difference of the cluster's value from the point of view of the question and the cluster's value from the point of view the answer.
Values are computed using term-frequency occurrences
Color coding is given in the right column
(For algorithmic details see Appendix 2.)

The darker the color, the better the match is. It can be seen the hypernyms barely make any improvement. There are three notes to be made here:

1. Differences depend on the clustering method
2. Other figures (not shown here) indicate that the FAQ-answer method is not efficient for representing (indexing) the clusters.
3. Point 2. may be spoiled by
    a. the small number of FAQs
    b. the hypernyms of the Word Associations, which does not reflect properly the relevant words and the relevant hypernyms for the "General" database (i.e., for database "Center for devices and radiological health")

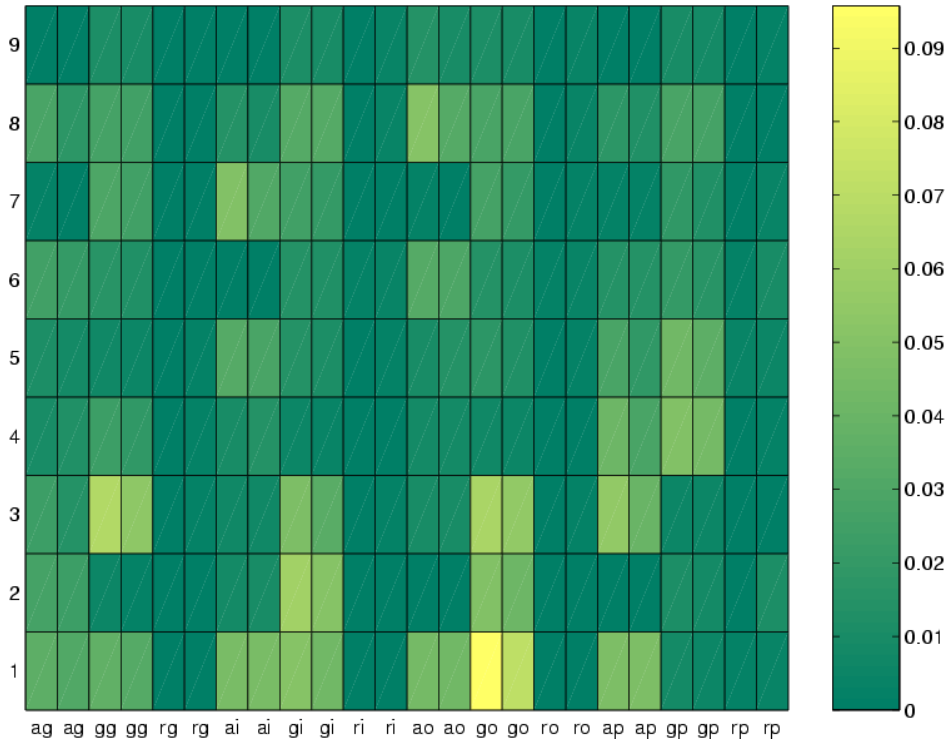

Figure 2. Evaluations of the FAQ—answer method

Database: All three databases. First letters of the horizontal axis: a=AHA, g=General, r=Radiology
Results are averaged over FAQs
Clustering method: All clustering methods: Second letters of the horizontal axis: g=greedy_outdeg, i=indeg, o=outdeg, p=PageRank
Column belonging to the first occurrence of the FAQ number: questions *without* hypernym augmenting
Column belonging to the second occurrence of the FAQ number: questions *with* hypernym augmenting
Vertical axis: cluster numbers (the smaller the number the smaller the size of the cluster)
Colors represent the squared difference of the cluster's value from the point of view of the question and the cluster's value from the point of view the answer.
Values are computed using term-frequency occurrences
Color coding is given in the right column
Threshold value: 0.7

Again, the darker the color, the better the match is. It can be seen the hypernyms make in improvement in few cases. The same notes apply as for Figure 1. We shall investigate the effect for the General database and for the "outdeg" (outgoing links) clustering method in the next Figure.

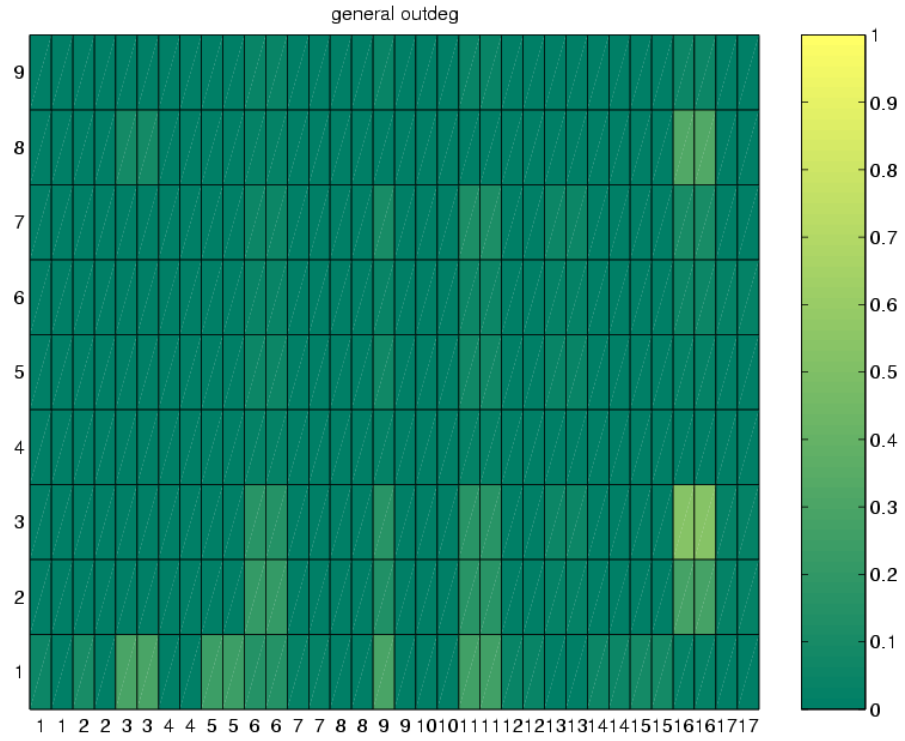

Figure 3. Evaluations of the FAQ—answer method

Database: General (Center for devices and radiological health)
Clustering method: outdeg
Horizontal axis: The number of the FAQ.
Column belonging to the first occurrence of the FAQ number: questions *without* hypernym augmenting
Column belonging to the second occurrence of the FAQ number: questions *with* hypernym augmenting
Vertical axis: cluster numbers (the smaller the number the smaller the size of the cluster)
Colors represent the squared difference of the cluster's value from the point of view of the question and the cluster's value from the point of view the answer.
Values are computed using term-frequency occurrences
Color coding is given in the right column
Threshold value: 0.7
(For algorithmic details see Appendix 2.)

According to Figure 3, the improvement noted in Figure 2 is mostly due to a single FAQ: FAQ number 9. In turn, the improvement using the hypernyms is limited. Notes concerning the quality of hypernyms are, however, valid. In particular, a few FAQs indicate that *FAQs could be representative* to clusters. This brute-force method alone, however, is not sufficient to index the clusters.

Results on keyword and key-phrase extraction methods

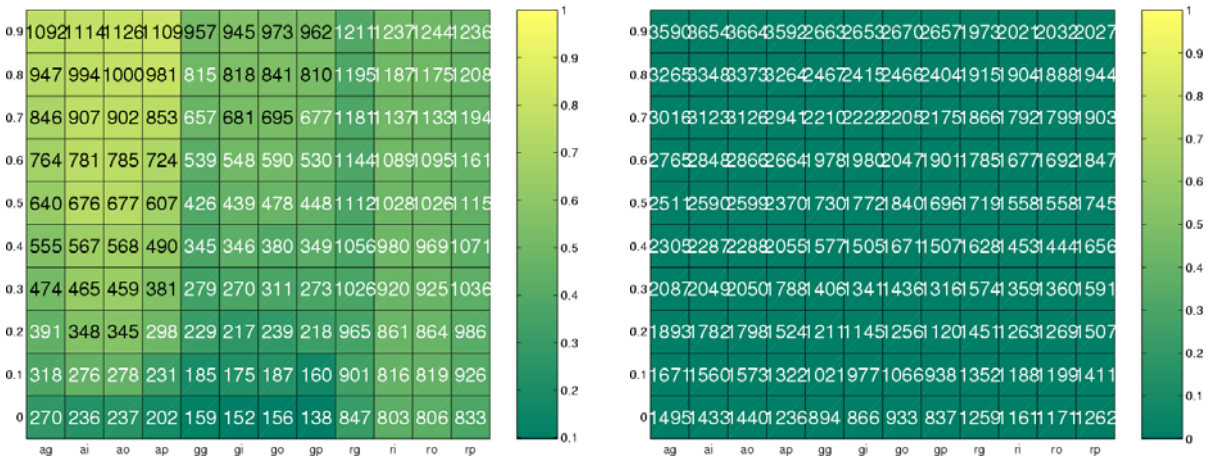

Figure 4. Keywords and key-phrases: *Methods and thresholds*
Left: keyword based on metatag, OR relation between keywords
Right: keywords from KEA, AND relation between keywords
Database: All three databases. First letters of the horizontal axis: a=AHA, g=General, r=Radiology
Clustering method: All clustering methods: Second letters of the horizontal axis: g=greedy_outdeg, i=indeg, o=outdeg, p=PageRank
Vertical axis: threshold values
Numbers within areas: number of keywords extracted at that threshold (horizontal), in that database (vertical first letter) and with that method (vertical second letter)
Color denotes averaged cluster values
(For algorithmic details see Appendix 2.)
*Keyword from metatags and method KEA provide similar results. Only one example for each are shown here.

Figure 4 shows that metatag and KEA based keywords and key-phrases are too many in number. Although in the OR[2] subfigure, values are high, in the AND[3] subfigure the values are low. That is, there is no (or very few) cluster(s), where the all the keywords or key-phrases of that cluster would appear in a single document. On the other hand, almost all documents contain at least one of the keywords or the key-phrases.

Four notes are due here:

1. Both KEA and the metatag methods are useful
2. They are not useful for the indexing of a large body of documents with simple AND or OR operations
3. They could be useful by using better AND, OR, XOR, etc. combinations of keywords and key-phrases. However, such indexing may give rise to a combinatorial blow-up.
4. Considerable differences can be seen between the different clustering methods.

[2] Keyphrase evaluation is based on disjunction of key-phrases: A document is 'good' if it contains one of the keyphrases.
[3] Keyphrase evaluation is based on conjunction of key-phrases: A document is 'good' if it contains all of the keyphrases

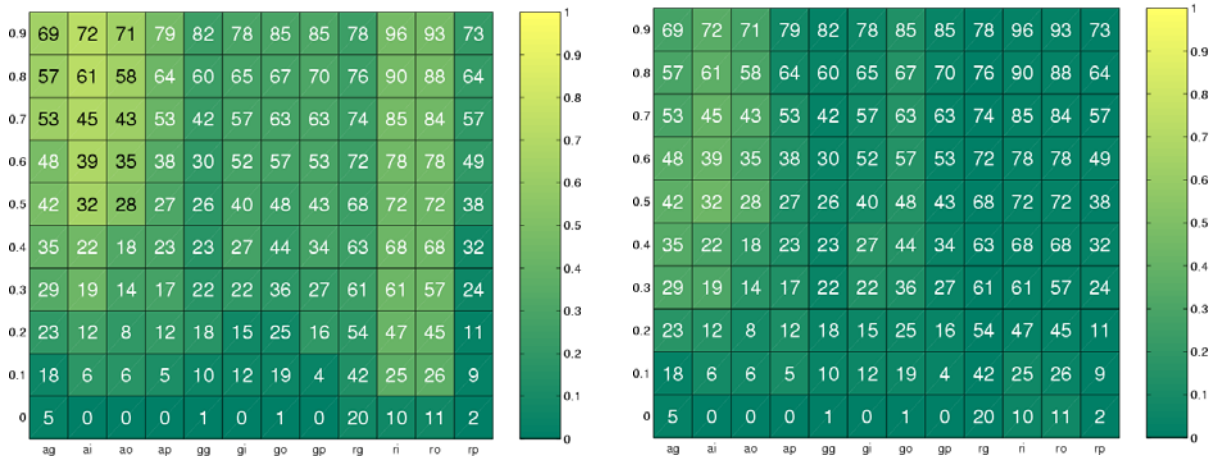

Figure 5. Keywords and key-phrases from Mutual Information: *Methods and thresholds*
Left: OR relation between keywords
Right: AND relation between keywords

Database: All three databases. First letters of the horizontal axis: a=AHA, g=General, r=Radiology
Clustering method: All clustering methods: Second letters of the horizontal axis: g=greedy_outdeg, i=indeg, o=outdeg, p=PageRank
Vertical axis: threshold values
Numbers within areas: number of keywords extracted at that threshold (horizontal), in that database (vertical first letter) and with that method (vertical second letter)
Color denotes averaged cluster values
(For algorithmic details see Appendix 2.)

Figure 5 shows that mutual information has very good chances to choose numbers, which can represent clusters. The task is to choose the appropriate threshold. The mutual information based keyword selection is somewhat better than the keyword selection based on the categories of the central pages of the databases (not shown here). This is promising, given that these databases are well established ones and the real task will concern not yet categorized clusters. In this respect, the keywords found be mutual information characterize (separate) the clusters somewhat better than the keywords chosen by the designers of the databases.

Notes are due here:

1. PageRank (4^{th}, 8^{th} and 12^{th} columns) is less characteristic than the other clustering methods
2. On database Radiology the AND relation is meaningless for all types of clustering methods. This is in accord with the observation that the OR relation on database Radiology is superior to all of the other databases. That is, database Radiology is (possibly) constructed by different principles.

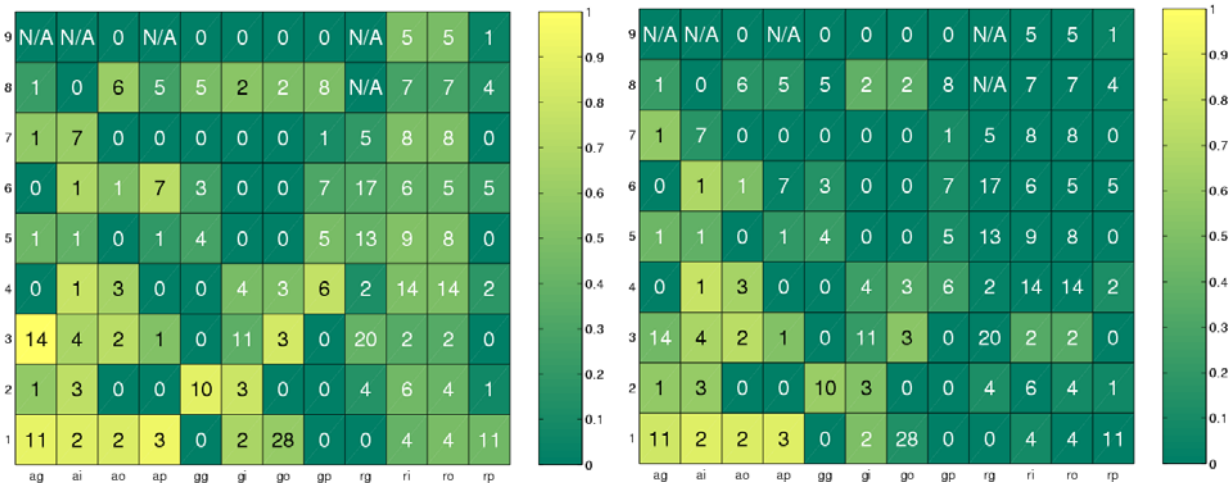

Figure 6. Keywords and key-phrases from Mutual Information: *Individual clusters*
Left: OR relation between keywords
Right: AND relation between keywords

Database: All three databases. First letters of the horizontal axis: a=AHA, g=General, r=Radiology
Clustering method: All clustering methods: Second letters of the horizontal axis: g=greedy_outdeg, i=indeg, o=outdeg, p=PageRank
Value of threshold: 0.3
Vertical axis: Cluster number. The smaller the number, the smaller the cluster is. (N/A denotes that some clustering methods produced less than 9 clusters.)
Numbers within areas: number of keywords extracted in that cluster (horizontal), in that database (vertical first letter) and with that method (vertical second letter).
Color denotes averaged cluster values
(For algorithmic details see Appendix 2.)

Figure 6 demonstrates that results of Figure 5 are relevant. Recall that in the FAQ-answer method, improvement was due to an improvement in a single cluster. Words extracted by the mutual information method is capable to represent and separate several clusters either in 'OR' or in 'AND' combinations. In turn, there are methods, which can produce indices, the first step for making robotic intelligence amenable for humans.