# Appendix B - Results from conceptualization

## Original document size: 289 words

The technique proposed is based on two interesting observations concerning the very nature of hyperspaces. Firstly, documents in a hyperspace don't need a predefined curriculum (like text-books do), but allow for linkage among documents. This in turn leads to a more natural decomposition of large documents into smaller ones that are linked together. Secondly, formatting a document normally involves the identification of important concepts of the text, and these normally get emphasized in some way by the author in order to improve readability. Now the importance of the HTML-tags should come clear: As noted above, HTML-files contain the formatting tags along with the text. Assuming that the author have structured and linked the documents (intelligently) while marking up important concepts within a document, we now have another (untraditional) way of extracting information from a document regarding its content. Different tags have different purposes and therefore is of varying importance. Therefore a predefined base of rules that tells what to do when encountering tags, yields both power and flexibility to the process of analysing HTML-documents. New rules may be easily added to the base without changing the implemented adaptive system if the base is explicit to the system. An appealing strategy seems to be as follows: "parse the document until a useful tag is run into. Analyse the information within the tag based on the rules from the rulebase. Do the appropriate action as told by the rule (i.e. retag/conceptualize the information), and continue parsing. When the end of the document is reached, a list of concepts and relations should be the result. This list is the basis for constructing the semantic network. (Perhaps that list is the semantic network, if nodes are written to a prolog-knowledgebase as new concepts are encountered...?)

## When stopwords and lexical analysis is perfomed, 133 terms remain

technique proposed based observations concerning nature hyperspaces firstly documents hyperspace don predefined curriculum text books allow linkage documents leads natural decomposition documents ones linked secondly formatting document normally involves identification concepts text normally emphasized author improve readability importance html tags noted html files contain formatting tags text assuming author structured linked documents intelligently marking concepts document untraditional extracting information document regarding content tags purposes varying importance predefined base rules tells encountering tags yields power flexibility process analysing html documents rules easily added base changing implemented adaptive system base explicit system appealing strategy follows parse document useful tag run analyse information tag based rules rulebase appropriate action told rule retag conceptualize information continue parsing document reached list concepts relations result list basis constructing semantic network list semantic network nodes written prolog knowledgebase concepts encountered

## When stemming is applied, 133 words (46% of original size) remain

techniqu propos base observ concern natur hyperspac firstli docum hyperspac don predefin curriculum text book allow linkag docum lead natur decomposit docum on link secondli format docum normal involv identif concept text normal emphas author improv readabl import html tag note html file contain format tag text assum author structur link docum intellig mark concept docum untradit extract inform docum regard content tag purpos vary import predefin base rule tell encount tag yield power flexibl process analys html docum rule easili ad base chang implem adapt system base explicit system appeal strategi follow pars docum us tag run analys inform tag base rule rulebas appropri action told rule retag conceptu inform continu pars docum reach list concept relat result list basi construct semant network list semant network node written prolog knowledgebas concept encount

## Removing duplicates further improves slightly: 87 words
## (30% of the original size)

```
action ad adapt allow analys appeal appropri assum author base basi book chang concept
conceptu concern construct contain content continu curriculum decomposit docum don easili
emphas encount explicit extract file firstli flexibl follow format html hyperspac identif
implem import improv inform intellig involv knowledgebas lead link linkag list mark natur
network node normal note observ on pars power predefin process prolog propos purpos reach
readabl regard relat result retag rule rulebas run secondli semant strategi structur
system tag techniqu tell text told untradit us vary written yield
```

## Example of Report:

## Document candidates of file4.html

```
debat:  252
mae:  231
user:  109
interfac:  105
adapt:  103
domain:  103
intellig:  101
model:  101
:  51
patti:  34
ben:  33
schneiderman:  31
speech:  4
futur:  3
direct:  3
manipul:  3
design:  2
system:  2
memori:  2
solv:  2
focu:  2
focus:  2
difficult:  2
anthropomorph:  1
comput:  1
move:  1
live:  1
entiti:  1
screen:  1
vision:  1
collabor:  1
filter:  1
appear:  1
predict:  1
lead:  1
unpredict:  1
feel:  1
job:  1
magic:  1
word:  1
smart:  1
mislead:  1
leav:  1
avail:  1
nl:  1
short:  1
term:  1
degrad:  1
```

```
level:  1
perform:  1
come:  1
issu:  1
critic:  1
time:  1
restrict:  1
avoid:  1
mistak:  1
essenc:  1
simpl:  1
blind:  1
peopl:  1
spatial:  1
process:  1
litteratur:  1
below:  1
nice:  1
possibli:  1
distinguish:  1
disagr:  1
mainli:  1
due:  1
look:  1
structur:  1
task:  1
profession:  1
novic:  1
dynam:  1
agre:  1
lot:  1
ambigu:  1
approach:  1
correct:  1
complex:  1
increas:  1
deleg:  1
agent:  -90
featur:  -198
folei:  -198
accord:  -199
softwar:  -199
cubricon:  -199
multilingu:  -199
**********************************
```

## Element candidates of file4.html#elm0 of elementtype: h1

```
debat:  -399
:  -399
-----------------
```

## Element candidates of file4.html#elm1 of elementtype: ul

```
user:  106
agent:  105
interfac:  104
adapt:  103
intellig:  101
model:  101
futur:  3
design:  2
system:  2
speech:  2
memori:  2
direct:  2
manipul:  2
folei:  2
anthropomorph:  1
```

```
comput:  1
patti:  1
move:  1
live:  1
entiti:  1
screen:  1
vision:  1
collabor:  1
filter:  1
featur:  1
appear:  1
predict:  1
lead:  1
unpredict:  1
feel:  1
job:  1
magic:  1
word:  1
smart:  1
mislead:  1
leav:  1
avail:  1
nl:  1
short:  1
term:  1
degrad:  1
level:  1
perform:  1
solv:  1
come:  1
issu:  1
critic:  1
time:  1
restrict:  1
avoid:  1
mistak:  1
essenc:  1
simpl:  1
accord:  1
blind:  1
peopl:  1
spatial:  1
process:  1
litteratur:  1
focu:  1
:  -399
-----------------
```

## Element candidates of file4.html#elm2 of elementtype: ul

```
agent:  105
user:  103
domain:  103
interfac:  101
focus:  2
speech:  2
difficult:  2
below:  1
nice:  1
possibli:  1
direct:  1
manipul:  1
distinguish:  1
softwar:  1
disagr:  1
mainli:  1
due:  1
ben:  1
look:  1
```

```
structur:  1
task:  1
profession:  1
focu:  1
novic:  1
dynam:  1
agre:  1
lot:  1
ambigu:  1
solv:  1
approach:  1
cubricon:  1
multilingu:  1
featur:  1
correct:  1
complex:  1
increas:  1
deleg:  1
:  -399
-----------------
```