Summer 6-2012

Automatic Semantic Trend Analysis of the Bled eConference: 2001-2011

Heinz Dreher
Curtin University, Australia, h.dreher@curtin.edu.au

25th Bled eConference
Special Issue
June 17 - 20, 2012; Bled, Slovenia

Automatic Semantic Trend Analysis of the Bled eConference: 2001-2011

Heinz Dreher
Curtin University, Australia
h.dreher@curtin.edu.au

Abstract

Semantic analysis algorithms have developed over the last decade to the point where they are almost within reach of everyone, as is Google for text searching. This study reports on an experimental application of automated semantic analysis to the Bled eConference 2001-2011 proceedings full text corpus. Rubrico, the specific tool used in the study is introduced. The methodology used to deploy Rubrico on the Bled corpus for the purpose of revealing the embedded concepts is explained. Interpretation and discussion are offered to indicate the possibilities ensuing from the semantic analysis. Further and future work is indicated to address limitations and further explore the prospects.

Keywords: conceptual analysis, concept trend analysis, semantics, Bled eConference, full text corpus.

1 Introduction

Semantic analysis of textual material is concerned with the extraction or identification of groups of terms with related meaning. These groups form high-level concepts or conceptual themes. Human readers do this automatically as they make sense of the documents they read and process. A good test of semantic analysis for a human would be to request the creation of an abstract based on a document. To the extent that the abstract represents the important or key concepts being dealt with in the document, one may make a judgement as to its correctness or usefulness. A major limitation of human semantic analysis is that it is labour-intensive and requires considerable time. This has been one of the motivations for research on automated semantic analysis. Nowadays, the large scale and general availability of text documents through digital libraries and other published corpora provides opportunity for scaling up the semantic analysis process from that which human readers can do, through semi-automated methods, to fully automated conceptual analysis of vast repositories of textual documents.

In our work on Automated Essay Grading (Williams and Dreher, 2004), which analyses student assignments and provides a grade and feedback based on the level of treatment of the concepts called for in the 'model answers', we realised the potential for addressing related problems such as plagiarism checking (Dreher, 2007), and to

improve Web search through automatic discovery of the user's conceptual model (Zhu and Dreher, 2010).

There are many other examples of the application of automated semantic analysis, for example in so-called recommender systems and trend analysis systems. These are information filtering systems designed to analyse user preferences at political party election time (Scharl and Weichselbraun, 2008) or for discovering consumer behaviour trends (e.g. WebLyzard, 2012).

Automated semantic analysis systems rely on prior research in the areas of machine learning, clustering, categorisation, have their roots in the information retrieval work of Gerard Salton first published in the 1960s (Salton, 1968), and deploy combinations of mathematical algorithms from these domains for the specific intended purposes.

During 2009 we embarked on a project to create a software tool that we named Rubrico (Reiterer et al. 2010) to allow the user to select suitable well-established statistical analysis algorithms used in the computational linguistics and information retrieval communities and combine their power in application to a given corpus.

Since 2001 the Bled eConference Proceedings have been digitally available as full text, and are therefore amenable to computational analysis. For the years 1995 to 2000, only the abstracts of the papers are available, making an automated concept analysis less feasible or interesting. For the 25th eBled eConference there was an opportunity to contribute a semantic analysis of the published papers, and this is the objective of the study reported here.

2 Objectives

Research dissemination events such as conferences and scholarly journal issues are normally centred on particular themes or disciplines chosen by the organisers and editors. Since the Information Systems discipline is relatively young and characterised by rapid new development of sub-disciplines driven mainly by advances in technology, it is relatively rare to find a conference series that has existed for a quarter of a century.

The Bled conference has been a long running "thematic conference series in or associated with the IS discipline" (Clarke, 2012) and in 2012 celebrates its 25th year of continuous operation. It would seem fitting therefore to discover via a thematic analysis just what these themes have been, and how they have changed over the years.

Our work here is to report on a study that explores our attempts at automated discovery of some underlying trends and patterns in a large conference database, focussing on the Bled conferences 2001-2011, for which the full text is available. Specifically, we conduct an automated semantic analysis (via *Rubrico*) of the Bled 2001-2011 Conference Proceedings corpus, and attempt to derive some insight from that analysis into the thematic trends latent in the published material.

3 Rubrico analysis methodology

Rubrico exists as a prototype and has been trialed in limited settings only. The process used to deploy *Rubrico* is given in Figure 1. Relying substantially on the prior work of Cimiano and Völker (2005) who developed Text2Onto, *Rubrico* provides a workflow and visualisation interface to help the user manage the analysis process. In addition, the

user may manually edit an automatically derived ontology. For readers unfamiliar with the concept of ontology as used in modern Information Systems, the term conceptual-structure, thematic-structure, or taxonomic relation may be used.

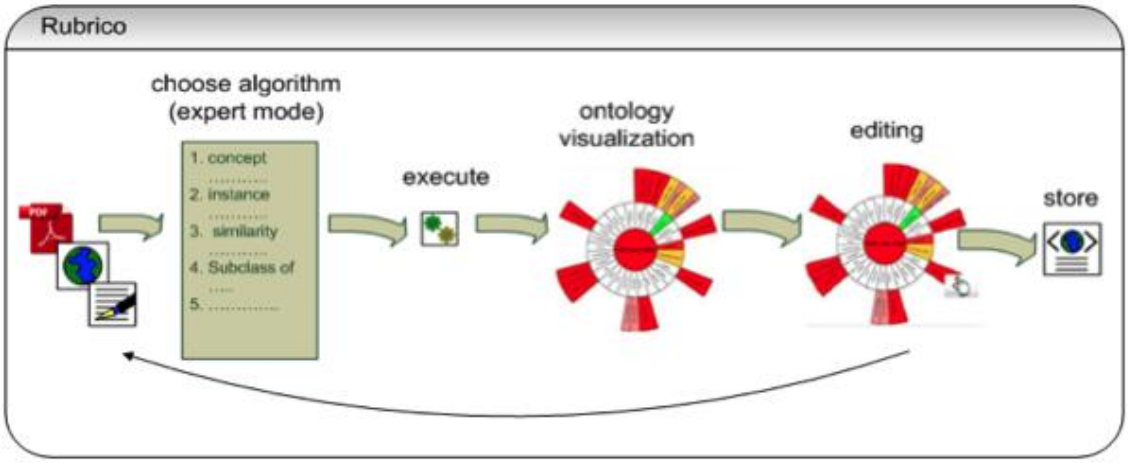


Figure 1: Rubrico Process (source: Reiterer, Dreher, & Gütl, 2010)

In *Rubrico*, the algorithms that learn taxonomic relations are grouped according to their purpose for extraction of *concepts*, *instances*, *similarity*, *subclassOf*, i*nstanceOf*, *relations*, and *disjointClasses*. These 7 categories of algorithm were derived from the literature as being potentially useful to our need, however in this analysis we have used algorithms from the *concepts*, *subclassOf*, and i*nstanceOf*, categories only, as these are the ones fully implemented in our prototype. *Rubrico* is currently still in development and this is the first large-scale case study we have applied it to. Thus we are just beginning to understand its power and its limitations, and to verify its results. In fact, this is the first study that has a parallel human powered analysis (Clarke, 2012) against which to compare the automatically computed result, although this must be left to a subsequent study as the results are not yet published.

A typical user view of the tool is given in Appendix 1. The top left panel shows the selected statistical-linguistic feature analysis algorithms applied. Of the 17 possible algorithms in the process of being implemented, the selection comprises just five (shown by the green dots). The bottom left panel shows the currently selected part of the corpus (for this test-run, the Bled_2001-2011_Abstracts – also displayed in the right panel, as a caption in the centre of the red circle).

Rubrico computes hierarchies of concepts. For each selected corpus, the computed hierarchy is shown as a graphical visual representation using the radial space-filling tree (Collins et al. 2009) as in Appendix 1, and in list form, in Figure 2.

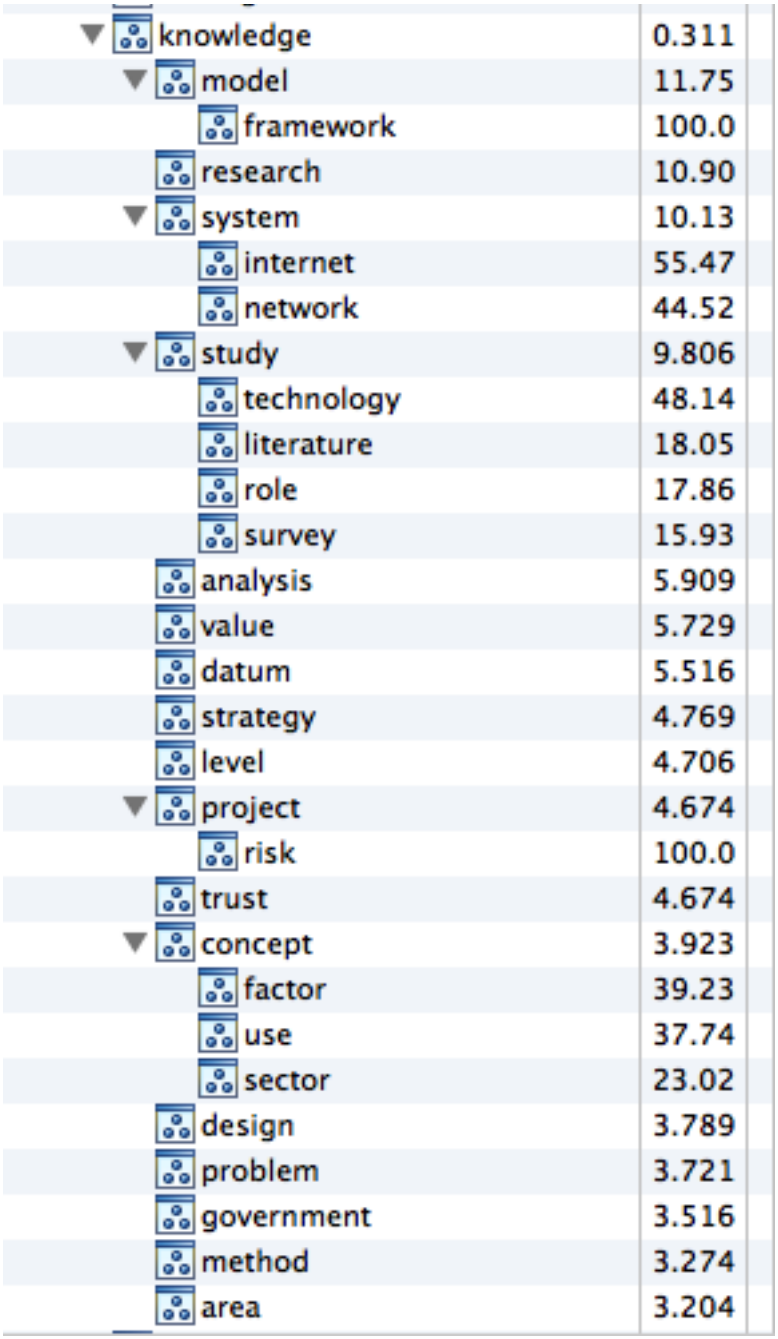


Figure 2: Ontological structure of concept "knowledge" computed from Abstracts

As an example of a concept hierarchy (3 levels deep), consider the top of Figure 2. The concept "knowledge" subsumes "model", which itself subsumes "framework". Typically, concepts are identified via a thesaurus, reference ontology, and word taxonomies such as that implemented in the lexical database WordNet for example (http://wordnet.princeton.edu/).

To quantify the 'importance' of terms in a document belonging to a corpus, various statistics can be used, and here as in Text2Onto, we use Term Frequency-Inverse Document Frequency (TF-IDF), and Relative Term Frequency (RTF) measures and

combine them into an average normalised score (in the range 0-100) called 'Rubrico-relevance' shown in the column at the right of Figure 2.

The first run of *Rubrico* on a corpus produces many hundreds of concepts that can be manually edited by selecting and deleting unwanted terms from the derived ontology. As currently implemented, this ontology-editing feature is inefficient. Despite this, it is useful to manually delete some frequently occurring terms that are of little interest to humans because doing so facilitates concentration on the remaining concepts. *Rubrico* may then be re-run with the human-edited parameters, resulting in a refined conceptual analysis. The computation time needed for a corpus of over 100 documents is at this stage excessively large, thus further constraining the practicality of numerous re-processing events.

After acquiring the Bled Corpus (2001-2011) from the conference organisers we investigated its parameters and compiled Table 1.

conference year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	total
#documents	*50*	*49*	*71*	*52*	*51*	*52*	*60*	*45*	*41*	*42*	*42*	***555***
ConceptKeywordCount	*4618*	*4229*	*5057*	*4461*	*4013*	*5041*	*5139*	*4776*	*4365*	*4341*	*4541*	

Table 1: #documents & Concept-keyword count for conference years 2001 - 2011

For each of the 11 Conference-years, a separate *Rubrico* analysis was done, delivering 11 ontologies of extracted concepts together with a "relevance" statistic computed by the above-mentioned algorithms as an indicator of importance of each *ConceptKeyword*.

4 Results

We have adopted the "relevance" statistic as an indicator of importance of each *ConceptKeyword*. That is, the higher the *Rubrico-relevance* (*Rr*) factors the greater importance the concept/keyword has to our consideration. Actually, it may be that the very low valued factors, or wildly varying factors, or trending factors, point to interesting events to follow up, but in this analysis we have focussed mainly on the high-valued factors.

Figure 3 depicts a fragment of the result of a *Rubrico* analysis for the first and last of the 11 years of conference proceedings in the Bled 2001-2011 corpus. It gives just a sample of the concepts retrieved as represented by the *ConceptKeywords*. Column 1 gives an identifying number, followed by two columns for each of the conference years 2001 and 2011. As shown in Table 1, the total number of *ConceptKeywords* is in the thousands, and varies from year to year.

In Figure 3, the columns to the left of the concept names are the *Rubrico-relevance* (*Rr*) factors and, over all of the concepts retrieved, they sum to 100, i.e. they are percentage values. The absolute value of *Rr* is not important; it is the relative values that can give an indication of any trend associated with a concept over a time dimension, or any other chosen dimension, a matter to be explored in the Interpretation and Discussion section.

	2001	2011
1	0.67 user	0.71 user
2	0.58 site	0.45 pp
3	0.55 com	0.41 relationship
4	0.51 transaction	0.41 student
5	0.51 ecommerce	0.35 group
6	0.48 pp	0.34 knowledge
7	0.40 goods	0.31 emotion
8	0.37 knowledge	0.30 website
9	0.36 implementation	0.30 interview
10	0.33 need	0.29 requirement
11	0.32 student	0.27 finding
12	0.32 supplier	0.27 actor
13	0.31 online	0.26 com
14	0.29 negotiation	0.26 implementation
15	0.28 relationship	0.24 respondent
16	0.27 environment	0.24 access
17	0.26 program	0.24 challenge
18	0.26 work	0.24 goal
19	0.26 government	0.24 cost
20	0.26 institution	0.24 participant
21	0.25 employee	0.24 transaction
22	0.25 industry	0.23 decision
23	0.25 category	0.23 communication
24	0.24 lack	0.22 country
25	0.24 html	0.22 experience
26	0.24 importance	0.22 impact
27	0.24 context	0.22 change
28	0.24 standard	0.22 nature
29	0.23 analysis	0.22 practitioner
30	0.23 criterion	0.21 control

Figure 3: Top30 Concepts in 2001 and 2011

From Figure 3, it can be readily seen that "user" featured strongly in 2001 (*Rr* = 0.67) and also in 2011 (*Rr* = 0.71). And concept "pp" also features in both years – but what is "pp"? We endeavour to explain some possible meanings for these in the next section.

Quite obviously, for the conference years where more papers were accepted one would tend to expect a greater number of retrieved concepts (e.g. the year 2007), however this is not a hard and fast rule as can be seen from year 2008 with 45 papers and 4776 concepts (*ConceptKeywordCount*) compared to year 2005 with 51 papers and just 4013 concepts – 6 papers more and 763 concepts less.

One of the most prevalent concepts in the Bled 2001-2011 corpus is represented by the term "user" (with two hidden nodes in the bottom right of Appendix 1) and comprises concept-keywords of "customer" and "consumer".

The concept represented by the term "knowledge" has 17 hidden nodes (top right Appendix 1) and has the ontological structure as shown in Figure 2. Of these 17 nodes, five have sub nodes: model subsumes framework; system subsumes internet, network;

study subsumes technology, literature, role, survey; project subsumes risk; and concept subsumes factor, use, sector. The column on the right in Figure 2 shows the relevance factor, again as a percentage. For example, since "framework" is the only contributor to "model", it represents a 100% contribution. The concept with name "study" has "technology" as the greatest contributor at 48.14% and "survey" with the smallest contribution at 15.93%. Whilst the "knowledge" hierarchy is only two deep, there are others in this analysis that are deeper. Theoretically there is no limit to the hierarchy depth, but practically it becomes less meaningful after 3 or 4 levels.

A process was devised whereby a first set of inferences could be drawn from the semantic analysis about changes in the importance of various concepts during the 2001-11 period. This process involved a sequence of steps which are presented in Figures 4, 4a and 4b, supported by Appendix 2, and which will be explained in the following paragraphs.

Firstly, the results for each year were sorted into descending order of *Rr*. Figure 3 shows the results for the first and last years of the set (i.e. 2001 and 2011), with the *Rr* for each concept in the 2nd and 4th columns. The top 30 for each of the 11 years were selected for further study.

Secondly, the top-30 concepts for all 11 years were merged into a single table, which therefore comprised 330 entries. That table was then sorted into three different sequences, and the results inspected. The purpose was to seek an appropriate basis for identifying relative importance among the *ConceptKeywords*. Figure 4a shows the top 30 arising from sorts based on three criteria, respectively relevance, keyword, and word occurrence count. Figure 4b shows the last 30 of the 330, for comparison. (In the 'B-year' column, B09-01 means Bled 2009, sequence-order 1 of 30, and B06-14 means Bled 2006, sequence-order 14 of 30).

The relevance column uses a derivative of the *Rr* measure that we call the *Rr_rank* (for *RubricoRelevance_Rank*). This is a number in the range 1 to *ConceptKeywordCount* as per Table 1, for each of the 11 conference years. It is used to simplify recognition of *ConceptKeywords* that may feature in further analysis. Thus, a *ConceptKeyword* with low valued *Rr_rank* has a relatively high *Rr* value; and a given *ConceptKeyword* may have a different *Rr_rank* over the 11 conference years - it is this fact that allows us to track the variation in occurrence.

Finally, a criterion was chosen, whereby a small sub-set of concepts could be isolated, to be tracked over all 11 conference years. We chose to focus on those concepts with the highest *Rr_rank* and appearing in the greatest number of Bled Conference years. This had the intended effect of being biased against short time-run concepts, and in favour of recurring themes. To these 30 were added seven concepts which the researcher considered provided useful counterpoints to those selected by statistical means.

Table 2 lists the resulting 37 concepts (*ConceptKeywords*). Appendix 2 shows the concepts over each of the 11 conference years (B_01, B_02 ... B_11) with matching *Rr_rank* and ordered in ascending rank (i.e. *Rr_rank* = 1 to some large integer as defined in Table 1 for each conference year).

note: top 30 (by relevance) for each year only

B-year	relevance	keyword	B-year	relevance	keyword	word occurrence count	B-year	relevance	keyword	word occurrence count
B09-01	1.118	pp	B10-05	0.375	year	2	B04-02	0.711	relationship	11
B04-01	1.076	pp	B05-09	0.316	year	2	B05-03	0.596	relationship	11
B08-01	1.051	pp	B01-18	0.261	work	2	B03-03	0.568	relationship	11
B07-01	0.907	pp	B07-19	0.244	work	2	B02-06	0.435	relationship	11
B06-01	0.895	user	B10-08	0.347	website	3	B11-03	0.414	relationship	11
B05-01	0.877	pp	B07-07	0.338	website	3	B09-05	0.372	relationship	11
B09-02	0.854	user	B11-08	0.301	website	3	B10-06	0.366	relationship	11
B06-02	0.724	pp	B10-17	0.254	view	1	B08-08	0.296	relationship	11
B08-02	0.716	user	B06-01	0.895	user	10	B06-08	0.287	relationship	11
B04-02	0.711	relationship	B09-02	0.854	user	10	B01-15	0.278	relationship	11
B11-01	0.71	user	B08-02	0.716	user	10	B07-11	0.272	relationship	11
B03-01	0.688	user	B11-01	0.71	user	10	B09-01	1.118	pp	11
B07-02	0.683	user	B03-01	0.688	user	10	B04-01	1.076	pp	11
B01-01	0.674	user	B07-02	0.683	user	10	B08-01	1.051	pp	11
B10-01	0.63	pp	B01-01	0.674	user	10	B07-01	0.907	pp	11
B03-02	0.62	pp	B05-02	0.619	user	10	B05-01	0.877	pp	11
B05-02	0.619	user	B04-03	0.518	user	10	B06-02	0.724	pp	11
B05-03	0.596	relationship	B02-01	0.516	user	10	B10-01	0.63	pp	11
B01-02	0.576	site	B10-16	0.254	understanding	1	B03-02	0.62	pp	11
B03-03	0.568	relationship	B02-18	0.261	type	1	B02-03	0.489	pp	11
B01-03	0.551	com	B09-30	0.246	trust	1	B01-06	0.484	pp	11
B10-02	0.523	people	B01-04	0.511	transaction	8	B11-02	0.453	pp	11
B04-03	0.518	user	B03-04	0.468	transaction	8	B10-04	0.4	knowledge	11
B02-01	0.516	user	B02-04	0.464	transaction	8	B01-08	0.374	knowledge	11
B01-04	0.511	transaction	B04-05	0.44	transaction	8	B04-09	0.362	knowledge	11
B09-03	0.509	http	B10-09	0.344	transaction	8	B11-06	0.34	knowledge	11
B02-02	0.508	supplier	B05-11	0.3	transaction	8	B03-07	0.338	knowledge	11
B01-05	0.505	ecommerce	B08-11	0.28	transaction	8	B02-08	0.326	knowledge	11
B04-04	0.497	supplier	B11-21	0.235	transaction	8	B08-07	0.3	knowledge	11
B02-03	0.489	pp	B09-24	0.254	tool	3	B09-12	0.292	knowledge	11
B01-06	0.484	pp	B08-27	0.243	tool	3	B05-15	0.267	knowledge	11
B03-04	0.468	transaction	B02-29	0.24	tool	3	B07-14	0.249	knowledge	11
B02-04	0.464	transaction	B09-23	0.254	theory	1	B06-17	0.231	knowledge	11

Figure 4a: Top 30 Concepts by Relevance, and # Years Occurring

note: top 30 (by relevance) for each year only

B-year	relevance	keyword	B-year	relevance	keyword	word occurrence count	B-year	relevance	keyword	word occurrence count
B06-14	0.233	design	B04-15	0.31	com	11	B01-25	0.238	html	1
B06-15	0.233	document	B05-10	0.301	com	11	B10-30	0.232	healthcare	1
B11-22	0.233	decision	B10-11	0.295	com	11	B10-03	0.405	health	1
B07-27	0.232	design	B11-13	0.264	com	11	B01-19	0.257	government	1
B10-29	0.232	element	B08-06	0.31	collaboration	2	B02-27	0.244	function	1
B10-30	0.232	healthcare	B07-09	0.282	collaboration	2	B03-25	0.241	form	1
B06-16	0.231	decision	B11-27	0.222	change	1	B08-28	0.241	firm	1
B06-17	0.231	knowledge	B11-17	0.24	challenge	1	B03-24	0.241	entity	1
B06-18	0.229	literature	B09-25	0.252	category	2	B11-07	0.305	emotion	1
B11-23	0.229	communication	B01-23	0.248	category	2	B04-10	0.357	emarketplace	1
B06-19	0.228	portal	B10-20	0.247	care	1	B09-19	0.266	effect	1
B06-20	0.228	structure	B10-23	0.244	behaviour	1	B04-23	0.251	dispute	1
B07-28	0.227	alignment	B08-17	0.249	basis	1	B10-21	0.247	disease	1
B07-29	0.227	interest	B06-13	0.233	author	1	B01-30	0.234	criterion	1
B07-30	0.225	document	B04-28	0.244	aspect	2	B11-19	0.238	cost	1
B06-21	0.224	actor	B10-22	0.244	aspect	2	B11-30	0.213	control	1
B06-22	0.224	employee	B01-29	0.234	analysis	1	B01-27	0.236	context	1
B06-23	0.224	investment	B07-28	0.227	alignment	1	B04-20	0.253	content	1
B06-24	0.224	message	B07-15	0.247	advantage	1	B02-25	0.249	consent	1
B11-24	0.224	country	B03-30	0.234	addition	1	B09-21	0.263	concept	1
B11-25	0.224	experience	B05-04	0.428	actor	8	B07-20	0.242	component	1
B11-26	0.224	impact	B07-03	0.396	actor	8	B04-18	0.259	commerce	1
B06-25	0.222	method	B09-04	0.394	actor	8	B11-27	0.222	change	1
B11-27	0.222	change	B02-12	0.282	actor	8	B11-17	0.24	challenge	1
B06-26	0.22	patient	B11-12	0.266	actor	8	B10-20	0.247	care	1
B11-28	0.22	nature	B04-25	0.248	actor	8	B10-23	0.244	behaviour	1
B06-27	0.218	ebusiness	B08-20	0.247	actor	8	B08-17	0.249	basis	1
B06-28	0.218	stage	B06-21	0.224	actor	8	B06-13	0.233	author	1
B06-29	0.218	survey	B05-20	0.252	access	4	B01-29	0.234	analysis	1
B06-30	0.216	interest	B11-16	0.24	access	4	B07-28	0.227	alignment	1
B11-29	0.215	practitioner	B10-26	0.237	access	4	B07-15	0.247	advantage	1
B11-30	0.213	control	B07-26	0.234	access	4	B03-30	0.234	addition	1

Figure 4b: Bottom 30 Concepts by Relevance, and # Years Occurring

Where "0" appears in a cell of Appendix 2, the meaning is that the corresponding concept (2nd column) did not feature in that conference year. For example, "group" did not appear in 2001. Note that it is the concept with name "group" that did not appear, and not necessarily the word, or string-of-characters forming the word, "group".

#	ConceptKeyword	#	ConceptKeyword
1	user	20	online
2	pp	21	finding
3	relationship	22	employee
4	knowledge	23	device
5	com	24	decision
6	group	25	access
7	transaction	26	http
8	implementation	27	people
9	communication	28	health
10	actor	29	ecommerce
11	supplier	30	year
12	requirement	31	emotion
13	participant	32	goods
14	need	33	website
15	student	34	resource
16	environment	35	researcher
17	structure	36	interview
18	site	37	finding
19	respondent		

Table 2: The 'Top' 37 *ConceptKeywords*

5 Interpretation and Discussion

With the automated semantic analysis (via *Rubrico*) of the Bled 2001-2011 Conference Proceedings corpus (first objective) achieved, we may now proceed to addressing the second objective of deriving some insight from that analysis. From Appendix 2, we have a list of 37 *ConceptKeywords* to form the basis of an 'interpretive discussion', through which some trends and perturbations that emerged from the conceptual analysis may be exposed. To assist with the discussion, the first 2 columns of Appendix 2 are reproduced as Table 2.

In each of the 11 conference years (2001-2011), the concept of ***user*** featured strongly, being ranked (*Rr_rank*) at either 1, 2, or 3, except for the year 2010 in which it achieved only 3834th place (row 1 in Appendix 2). This, at first, very surprising perturbation is easily explained. Consider Appendix 2, and note that in column with name "rank-B.10" (meaning the rank for the Bled conference year 2010) the *ConceptKeywords* ***people*** and ***health*** (rows 27 and 28) have values 2 and 3 respectively. This indicates that authors were using the term or concept ***people*** rather than ***user***, and the reader may now check that eHealth was a big feature of the 24th Bled eConference held in that year. ***People*** is another perspective on ***user***, and one may postulate that what we see here is the response by authors to the calls of the conference organisers and editors, adding weight to the proposition that editors have a big influence in the direction of the thinking of a

body of authors. To the extent that this is true one can verify that the semantic analysis (for example as per *Rubrico*) is creating a 'true' picture of reality.

The second most prominent *ConceptKeyword* to emerge is ***pp***. What an odd thing is that? ***pp*** is of course meaningless as a concept in the usual sense, however if we understand that the corpus is a collection of scholarly articles, for which the authors have created reference lists, often including a sequence of pages in their citations (indicated by "pp. 22-55", for example), then we may make some sense out of this. Is there a particular style of referencing being required which may explain the ***pp*** performance? Note that in the year 2001 the *Rr_rank* is 6, then climbing to 3, then 2, and very often at 1. This could, for example, be associated with an already-strong expectation of precise citation being tightened during the early years of the decade.

Relationship is the third *ConceptKeyword* identified in Table 2, and, as for ***knowledge*** (the fourth) its *Rr_rank* profile rises and falls but remains within the range of a low at 17 and a high of 2. Does such a consistent and strong performance indicate a very great emphasis, in this conference, on the pursuit of truth and explanation through systematic study and investigation of the essential connections between things? Readers may form their own view by referring to the ontology for these *ConceptKeywords* as depicted in Figure 5, Figure 2, and Appendix 2.

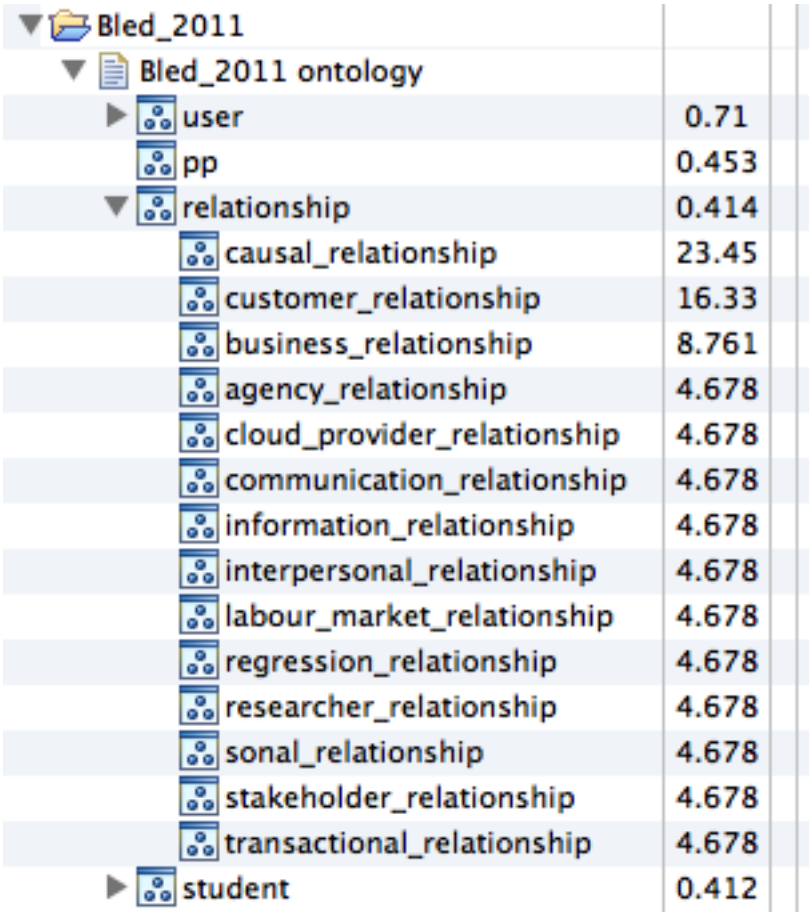


Figure 5: B_2011_relationship

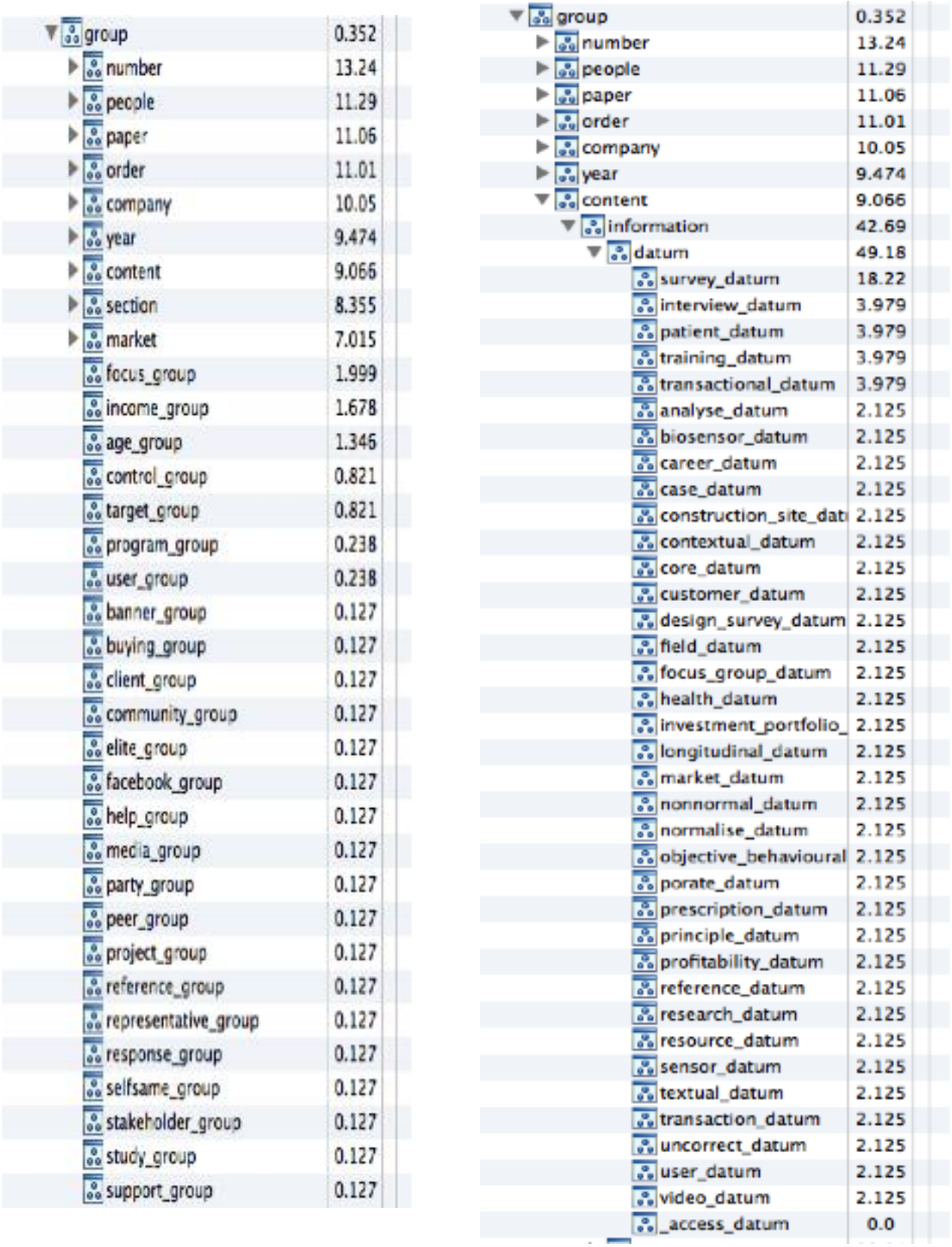


Figure 6: B_2011_group, and expansion of content hierachy

Item 5 featured in the 'top' 37 *ConceptKeywords* list is ***com***, which performs strongly over all of the conference years and is clearly associated with references to "dot com" and website URLs.

Group is the first item (according to a row-by-row consideration of Appendix 2) to have a zero score (in year 2001) then gradually, if a little erratically, growing in prominence over the ensuing decade. It may reveal an emergence of the role of people in teams and concern for societal issues in general.

In the left hand panel of Figure 6, one can see that group is rather an extensive structure, consisting of 34 elements, nine of which have sub-hierarchies. In the right hand panel of Figure 6, we see the ***content*** sub-hierarchy, which is 3-levels deep. Inspection of Appendix 2 for ***group*** reveals that this *ConceptKeyword* was absent from the 2001 proceedings.

Again, with reference to Appendix 2, we see that the following *ConceptKeywords* also have zero entries for one or more conference years: ***supplier*** (missing in 2003); ***participant*** (2010); ***need*** (02, 05, 08, 10, and 11); ***site***; ***respondent***; ***employee***; ***device***; and so on. *ConceptKeywords* ***access***, ***http***, ***people***, are remarkable because they appear only in 2005, then disappear for a period and perhaps reappear. This may be indicative of a fad, but would need much more in-depth exploration than has been possible here.

Continued analysis along the lines as offered above, and guided by some particular investigative purpose, or hypothesis, will serve purposes that heretofore could not be satisfied.

Clarke (2012) presents manual analyses of the Bled Conference corpus. The parallel development of that paper and this one has precluded formal comparisons being undertaken between them. It is striking, however, that the human-created ontology (in our terminology) is at a higher conceptual level than achieved by *Rubrico*. Combinations of terms such as "eMarkets, Directories, Auctions" are reported as being characteristic of the period 1998-2002 (Clarke 2012). For the 'super concept' formed from "eMarkets, Directories, Auctions" to be detected automatically it must have a textual association and eventual representation in the corpus. For the years 2001 and 2002 in our analysis the *ConceptKeywords* ***transaction*** and ***implementation*** may pertain; an intensive knowledge-elicitation and -engineering exercise would be needed to match this against the mentioned 'super concept'. Such analysis must await future attention as it is not within the scope of the current work.

6 Future work

As in all experimental research, there are limitations and deficiencies that one would like addressed. *Rubrico* makes possible the semantic treatment of vast amounts of text; but it is not intelligent. The human mind may find it difficult to comprehend certain ontological structures that *Rubrico* computes (e.g. ***group***). Therefore, an improvement that needs to be considered is for human editing of the initially-computed ontology, followed by a re-run of the semantic analysis.

Another limitation at present is the performance of *Rubrico* with large document sets – which we currently estimate as being greater than about 100 conference papers. Our initial attempt at analysis was to deploy *Rubrico* on a laptop computer, to deal with the entire 555 documents in the Bled 2001-2011 Conference corpus; it resulted in

'stagnation'. This points to a clear need for implementation in a more computationally-powerful environment.

Next, we want further functionality to automate the construction of Appendix 2 for example, and interactivity, interoperation, and dynamic visualisation of the elements of information structures depicted in the foregoing description and explanation. There is much work to do.

Despite the extensive wish-list indicated here, and the associated limitations, significant advances can be made by interested and enthusiastic researchers applying *Rubrico* (in whatever version it is or may become available) or similar semantic analysis tools, in the pursuit of insight not possible with the unaided human brain.

In order to check the usefulness of automated conceptual analysis of full text corpora such as has been attempted here, one would ideally need to engage in a comparison with the results of other types of analyses, and especially human-generated ones. As there is a special section of this 25th anniversary of the Bled conference, any alternate analyses published could form an interesting and useful agenda for further research.

Acknowledgement

The author acknowledges the following contributions: Emanuel Reiterer who developed the *Rubrico* concept analysis system during a masters project in 2009; Gregor Lenart who was instrumental in providing access to the full text of the proceedings for the years 2001-2012; Roger Clarke who encouraged the computer analysis of the full-text documents at the outset and assisted in obtaining access to the textual repository referred to here as the Bled 2001-2011 Conference Proceedings corpus. Especial thanks to the reviewers of the original draft of the paper – very helpful ideas and suggestions have been included as a result.

References

Cimiano P and Völker J (2005) Text2Onto – A Framework for Ontology Learning and Data-driven Change Discovery. Institute AIFB, University of Karlsruhe.

Clarke R. (2012) The First 25 Years of the Bled eConference: Themes and Impacts. Proc. 25th Bled eConference, June 2012

Collins, C, Carpendale, S and Penn, G (2009) DocuBurst: Visualizing Document Content using Language Structure. Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis '09), 28(3): pp 1039-1046.

Dreher, H (2007) Automatic Conceptual Analysis for Plagiarism Detection. Journal of Issues in Informing Science and Information Technology (IISIT), Vol 4, pp601-614.

Reiterer, E and Dreher, H and Gütl, C (2010) Automatic Concept Retrieval with Rubrico In Schumann, M, Kolbe, LM, Breitner, MH and Frerichs, A (ed), Multikonferenz Wirtschaftsinformatik - MKWI 2010, Feb 23 2010, pp3-14. Göttingen, Germany: Universitätsverlag Göttingen.

Salton, G (1968). Automatic Information Organization and Retrieval. McGraw Hill Text.

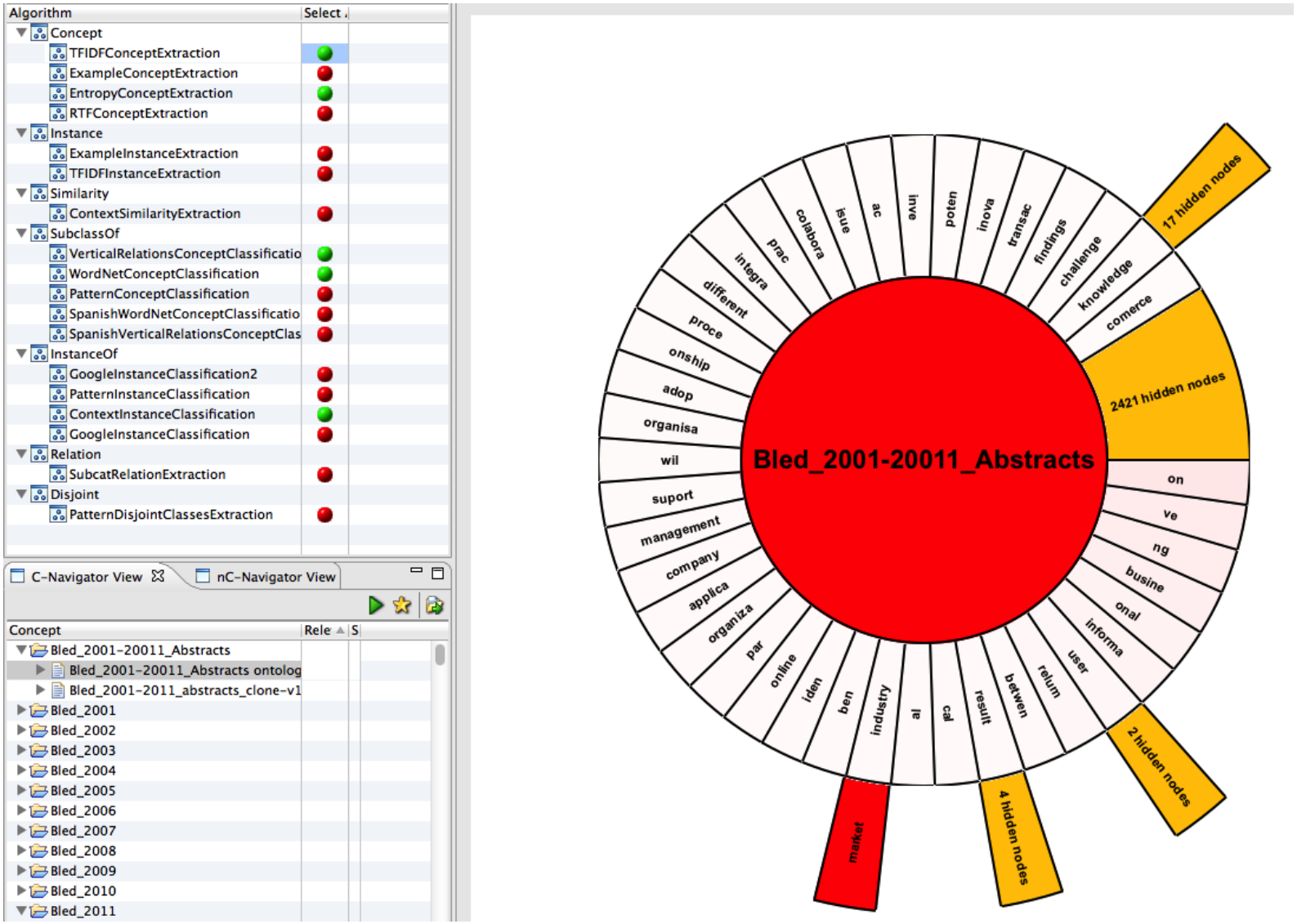
Scharl, A and Weichselbraun, A (2008) An Automated Approach to Investigating the Online Media Coverage of US Presidential Elections, Journal of Information Technology & Politics, 5(1): 1, pp21-132.

WebLyzard (2012) http://www.weblyzard.com Accessed 20th March 2012.

Williams, R and Dreher, H (2004) Automatically Grading Essays with Markit© Journal of Issues in Informing Science and Information Technology Vol. 1, pp693-700.

Zhu, D and Dreher, H (2010) Exploring semantic characteristics of socially constructed knowledge repository to optimize Web search. In HO Nigro & SEG Cisaro (eds), Ontologies Driven Web Mining: Concepts and Techniques, IGI Global, Hershey.

Appendix 1: Rubrico Features



Appendix 2: 37 ConceptKeywords and Their Usage over the 11 Conference Years

word order	keyword	B_01	rank-B.01	B_02	rank-B.02	B_03	rank-B.03	B_04	rank-B.04	B_05	rank-B.05	B_06	rank-B.06	B_07	rank-B.07	B_08	rank-B.08	B_09	rank-B.09	B_10	rank-B.10	B_11	rank-B.11
1	user	0.67	1	0.52	1	0.69	1	0.52	3	0.62	2	0.90	1	0.68	2	0.72	2	0.85	2	0.01	3834	0.71	1
2	pp	0.48	6	0.49	3	0.62	2	1.08	1	0.88	1	0.72	2	0.91	1	1.05	1	1.12	1	0.63	1	0.45	2
3	relationship	0.28	15	0.44	6	0.57	3	0.71	2	0.60	3	0.29	8	0.27	11	0.30	8	0.37	5	0.37	6	0.41	3
4	knowledge	0.37	8	0.33	8	0.34	7	0.36	9	0.27	15	0.23	17	0.25	14	0.30	7	0.29	12	0.40	4	0.34	6
5	com	0.55	3	0.46	5	0.37	6	0.31	15	0.30	10	0.40	4	0.36	5	0.34	4	0.36	6	0.30	13	0.26	13
6	group	0.00		0.29	11	0.39	5	0.32	14	0.21	63	0.35	5	0.34	6	0.35	3	0.28	15	0.25	18	0.35	5
7	transcaction	0.51	4	0.46	4	0.47	4	0.44	5	0.30	11	0.21	41	0.18	91	0.28	11	0.02	1300	0.34	9	0.24	21
8	implementation	0.36	9	0.22	43	0.24	27	0.25	24	0.32	8	0.21	36	0.27	10	0.34	5	0.22	50	0.27	15	0.26	14
9	communication	0.22	40	0.29	10	0.31	9	0.37	7	0.35	7	0.31	7	0.22	40	0.29	10	0.22	52	0.35	7	0.23	23
10	actor	0.09	280	0.28	12	0.19	79	0.25	25	0.43	4	0.22	21	0.40	3	0.25	20	0.39	4	0.12	197	0.27	12
11	supplier	0.32	12	0.51	2	0.00		0.50	4	0.37	6	0.25	11	0.36	4	0.17	99	0.17	98	0.10	261	0.19	62
12	requirement	0.18	84	0.28	13	0.17	108	0.14	170	0.26	17	0.20	49	0.24	25	0.26	12	0.19	75	0.28	14	0.29	10
13	participant	0.14	155	0.19	83	0.22	41	0.22	50	0.41	5	0.26	9	0.27	12	0.25	18	0.29	13	0.00		0.24	20
14	need	0.33	10	0.00		0.26	14	0.34	11	0.00		0.42	3	0.33	8	0.00		0.25	29	0.00	0	0.00	0
15	student	0.32	11	0.13	172	0.24	21	0.17	110	0.28	14	0.15	127	0.25	18	0.21	56	0.20	66	0.09	294	0.41	4
16	environment	0.27	16	0.00		0.23	31	0.30	16	0.26	16	0.17	91	0.00		0.26	13	0.35	7	0.22	41	0.00	
17	structure	0.21	52	0.23	41	0.20	65	0.25	22	0.16	138	0.23	20	0.17	98	0.26	14	0.26	22	0.20	62	0.00	
18	site	0.58	2	0.00		0.26	13	0.39	6	0.00		0.00		0.21	45	0.17	86	0.34	9	0.12	211	0.10	271
19	respondent	0.00		0.16	134	0.27	11	0.14	157	0.14	175	0.21	40	0.22	42	0.29	9	0.33	10	0.23	31	0.24	15
20	online	0.31	13	0.18	104	0.30	10	0.37	8	0.18	103	0.15	124	0.19	77	0.12	205	0.16	114	0.24	24	0.16	97
21	finding	0.16	104	0.16	126	0.23	33	0.20	69	0.24	35	0.21	45	0.22	38	0.25	16	0.27	17	0.29	12	0.27	11
22	employee	0.25	21	0.26	23	0.24	22	0.22	48	0.23	41	0.22	22	0.18	88	0.20	64	0.21	64	0.00		0.19	48
23	device	0.04	595	0.26	22	0.25	17	0.12	199	0.25	25	0.00		0.18	87	0.00		0.27	18	0.19	69	0.21	33
24	decision	0.18	72	0.20	66	0.21	52	0.26	19	0.20	72	0.23	16	0.24	21	0.00		0.00		0.22	39	0.23	22
25	access	0.00		0.00		0.00		0.00		0.25	20			0.23	26	0.21	55	0.19	72	0.24	26	0.24	16
26	http	0.00		0.00		0.00		0.00		0.19	83	0.00		0.00		0.00		0.51	3	0.00		0.00	
27	people	0.00		0.00		0.00		0.00		0.28	12	0.00		0.00		0.00		0.00		0.52	2	0.00	
28	health	0.02	1062	0.02	1297	0.04	631	0.05	501	0.09	329	0.10	267	0.02	1133	0.11	210	0.10	230	0.41	3	0.16	90
29	ecommerce	0.51	5	0.20	69	0.19	81	0.32	13	0.09	337	0.13	159	0.04	623	0.09	306	0.08	342	0.02	1150	0.02	1076
30	year	0.00		0.00		0.00		0.00		0.32	9	0.00		0.00		0.00		0.00		0.38	5	0.00	
31	emotion	0.00	2816	0.01	1490	0.02	1146	0.00		0.01	1441	0.01	1632	0.01	1657	0.01	1828	0.01	1414	0.01	1375	0.31	7
32	goods	0.40	7	0.18	92	0.32	8	0.24	33	0.19	85	0.25	10	0.15	126	0.13	176	0.10	229	0.02	913	0.09	314
33	website	0.15	129	0.13	177	0.13	187	0.16	129	0.00		0.00		0.34	7	0.18	80	0.10	258	0.35	8	0.30	8
34	resource	0.14	148	0.20	67	0.20	62	0.00		0.00		0.34	6	0.00		0.22	53	0.24	35	0.23	35	0.17	80
35	researcher	0.10	269	0.21	60	0.23	38	0.23	36	0.21	67	0.17	85	0.19	64	0.15	127	0.34	8	0.15	147	0.19	58
36	interview	0.11	214	0.20	74	0.20	68	0.22	55	0.00		0.19	65	0.22	37	0.16	115	0.21	60	0.16	128	0.29	9
37	finding	0.16	104	0.16	126	0.23	33	0.20	69	0.24	35	0.21	45	0.22	38	0.25	16	0.27	17	0.29	12	0.27	11