Course validation using an automatic system for sorting similar technical papers.

N. Faure
Research Associate, team Modeme, IAE, Lyon 3

R.M. Faure , J.C. Gress
Professor, ENTPE, Vaulx en Velin, France

G. Hémond
Professor, EPFL , Lausanne, Switzerland

ABSTRACT: As part of the geotechnics course in ENTPE (National College of Public Works, France), the validation of course credits depends on the synopsis, by the student, of two or three recent technical papers (most of the time in English language). In order to ensure that the articles are up-to-date and from the most recent conferences, we have developed a knowledge base tool, which automatically recognises "similar papers" from a range of conferences. With these "similar papers" the student must compare, and perform a critical analysis of different points of view about the same topic.

This new methodology for course credit validation is appreciated by the students, who often are so enthusiastic that they submit work over and above that which is required. Less motivated students at least do the translation of the papers into French, thereby assimilating the concepts contained therein.

The entire validation process goes from the registration of students, how they download their papers, remain in contact with their teacher, submit their work and get their grade from a web-site.

The general framework of the tool that analyses all the technical papers and finds similar documents will be briefly explained and some examples shown. This tool is also quite useful for PhD students for managing large bibliographies.

1 INTRODUCTION

ENTPE (National College of Public Works, France), is a part of the French Ministry of Publics Works, for which it provides approximately 150 civil engineers each year. Civil engineering is, in fact, extended to transportation, sustainable building and environment. Using EC rules all courses provides ETCS credit to students who obtain an engineering degree after three years in this school and a sufficient number of credits.

2 TEACHING SOIL MECHANICS.

Students are selected by competition, after two years of preparatory classes. They spend three years in ENTPE (semesters 5 to 10) to achieve their engineering qualification. During these three years, soils mechanics forms a very important part of their civil engineering studies.

It is in the sixth semester that they begin to study soils and its properties. The fundamentals of soils mechanics is given to the students: soils components, seepage, shear strength mechanisms, settlement and consolidation. This course contains 8 lectures of 3 hours each, that is, 24 hours for 1 ETCS credit. During this semester an introduction to earth sciences: geology and sustainable development during 30 hours counts for 2 ETCS credits.

At the beginning of the seventh semester (second year in ENTPE) a very interesting course teaches students how useful the geological approach is for guiding investigations during a real project and detecting problems during the works. A wide range of cases are analysed. All the other teachers are practising engineers and they can transfer their practical skill and knowledge to the students. Like to previous courses, lectures are given to small groups of students (24 max) so the teaching staff contents 8 engineers under the supervision of a professor. One ETCS credit is due to this course, when validated.

The validation of these two courses, in order to obtain the credits, is a two hours examination, solving three or four small problems.

The course to which we refer in this paper is given during the eighth semester to half of the students who have already chosen a specialisation in civil engineering. It is devoted to soil mechanics works and consist of height lectures of three and half hours (28 hours). These lectures are: Investigations, shallow foundations, deep foundations, rigid screens like retaining walls, flexible screens like sheet piles or diaphragm walls, drainage and dewatering, rocks mechanics for slopes and tunnels, slopes and embankments. For each lectures the teacher is an engineer

working in the field, and the lecture is repeated to all groups. Consequently the time table is quite more complex, all students doing Applied Soils Mechanics (the title of this course) at the same time, but with different subjects depending of the group. The contribution of the teacher, who is usually working in a industry, his investment is very significant as he speaks about the domain in which he is an expert, and as his lecture is repeated, he can easily improve it, week by week. The student has the benefit of up to date information, skill and know how in the topic. During the lecture questions obtain the more accurate answers. After validation the student obtains 2 ETCS credits.

2.1 *Course validation system*

The number of hours devoted to this course is quite small compared to the overall subject. To ensure the best use of this time, we also use networks to extend the teaching. The procedure for validation follows these steps. During the lecture time, students have, by e-mail, to propose binomial to the professor, and a choice within the topics of the course. For each binomial the professor gives it links for several files (pdf files). These files are papers from conferences or symposium with a strong connection to the topic in question. The students download the files and must do a synthesis of the papers, sending it back as attached document to the professor. For this synthesis they have to translate the documents, (documents are mainly in English language for French students and in French language for foreign students), and discuss the differences or the similarities between the papers. The main difficulty for the professor is to provide to the students with similar papers including theories (seepage and drainage, earth pressure, soil behaviour…) and described cases (foundations, retaining structures, slopes embankments…) corresponding to the students request. The use of the home-built tool called MKD (as Managing Knowledge by Domain) is an appreciated aid.

3 BUILDING A KNOWLEDGE BASE

The Managing Knowledge by Domain (MKD) system is built for knowledge management (KM) using all the information contained in the proceedings of a conference or a set of conferences in the soils mechanics domain. The first step for building a knowledge base is to create a taxonomy of all the expressions used in the proceedings.

3.1 *Taxonomy*

The conceptual content of a paper is determined from its terminology. We have two lists of expressions: words from the domain and common words. Within domain terminology by comparing expressions from papers we can appreciate their likelihood.

First we must build the taxonomy of the domain. For this, all the papers of the symposium are treated like a single file by Split_plus. The process identifies expressions of the whole text. Expressions are defined by 6 patterns which are N, AN, NPN, NPAN, NPNA, and NPNPN.

N means name, A adjective, P preposition.

Table 1 summarizes this construction.

Patterns	Found expressions	Different expressions	Expressions in thesaurus	Most employed expressions
N	249622	4463	695	slope (9712), landslide , soil, rock, analysis
AN	48211	18561	332	Finite element (328), expansive soil (133)
NPN	16137	9538	393	Factor of safety (396), stability of slope (72)
NPAN	5435	4259	207	Angle of internal friction (23)
NPNPN	790	630	307	Effect of earthquake on dam (9)
Total	320195	37401	1934	

Table 1 : Number of expressions used when building the taxonomy from the 270 papers presented at Xi'an ISL congress

The « taxonomy » (kind of thesaurus) contains 1934 words or expressions with a strong link with geotechnics and slope stability.

The 1934 words or expressions in the taxonomy are sorted by an expert from the 320195 expressions detected in all the texts of the conference.

3.2 *References*

All listed papers from the symposium have a structure defined by the conference organizers with names of authors, title, abstract, body text and references that will be exploited to increase the knowledge data-base.

Pdf files of articles of the symposium ISL10 (270 files) were placed in a directory (taken from a CD) and are processed one by one to extract the title, authors (split_moins tool), abstract, references (super_split tool) and all expressions with text frequency (split_plus tool). (see below MKD flowchart)

These three tools provide three files that are processed by the user-friendly front end of MKD and all results of this process are stored in the

knowledge-base. Counters that are associated with each item allow a continuous survey of the work.

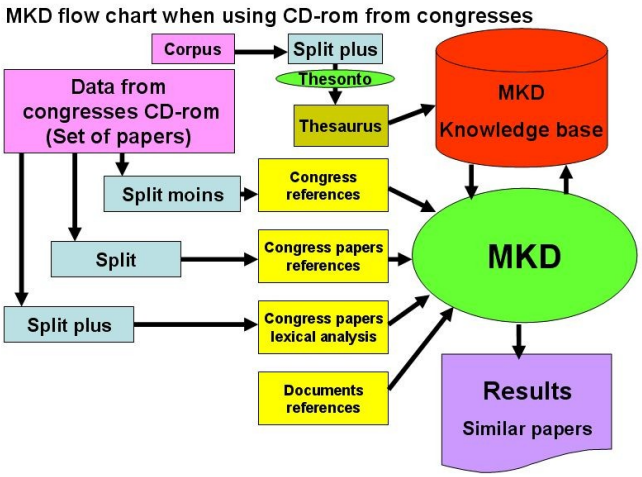
Split_moins gives the list of the 270 papers, and when feeding the data base, titles, authors and addresses of authors are recognized and stored in the data base.

Super_split extracts from each paper the abstract and the bibliography. This bibliography is compounded of references that illustrate the processed paper; these references are stored in the data-base keeping a link between them and the paper from which they are obtained.

With the addition of references of analyzed bibliographies in each paper and the volume of the knowledge-base increases very rapidly.

Split_plus analyses the text of each paper finding all the expressions in each text. The same process was used for taxonomy.

For each paper, a "signature" terminology is derived from the results of split_plus used file by file. We call “signature” of a text, the list of the expressions and their occurrences, found inside the text and also in the “thesaurus” already built. The “signature” of each text is between 100 and 600 expressions. All is stored in the knowledge-base.



4 HOW THE BASE CAN BE QUERIED?

Knowledge from the references in articles is important and gives a picture of the domain. The usual questions of the base will therefore provide lists of references, which will serve in a second step for an approach to the content.

4.1 *Checking*

Spell checking of proper names (people and places) is performed from lists of references dynamically built from the elements already in the knowledge-base. This validation is done after displaying the few new words encountered at each pass. Similarly, lists of synonyms can easily recognize the different forms of phrases. MKD includes numerous sequences of controls and verification to ensure the uniqueness of the recorded information. If references to articles from the reading of the article are reliable, those written by the authors in the bibliographies of articles are less, varying typographic or syntax. A search for similar references is nevertheless possible and the rare ambiguities are removed by the user. For journals, the name used is maintained and is associated with a more generic name in order to address the processing of bibliometrics.

4.2 *Obtaining lists of items (sorting according to criteria or bibliometrics)*

From sorting (questions are predefined in SQL language) on the author names, years, journals, references of each article you get the answers to these questions:

What papers from this (these) author (s)?

Who are the authors that use the word (s) in their title?

Which are the most cited articles?

Which are the articles (or authors) that use such expression(s)?

Which are the articles that cite this reference? etc..

A list of references can be the sum of the result of several questions; the list goes to each question with elimination of duplicates.

The use of these lists allows bibliometrics.

Who are the authors of the articles from this list? (Identification of a community)

Which are the articles that refer to this list of words? (Extension of the community)

What are the expressions and occurrences of this list? (Conceptual content of a list of articles)

Which are the articles that use such concepts? (Search by content)

The use of conceptual content starts from simple comparison of two conceptual contents (search new expressions) to more elaborate approaches that are, using the frequencies associated with each expression, vector approach, log-likelihood, tf-idf calculations allow more meaningful conclusions.

The expressions that emerge from a calculation of log-likelihood (or tf-idf) are the specific phrases from the body of reference - which means that if a document does nothing specific in relation to the corpus, it will present only low values (apart from a few marginal changes due to stylistic differences). (Duning, 1993) (Beney, 2008)

It is these last possibilities that allows the sorting of papers by likelihood. We extract from the knowledge-base the signature of a paper and MKD computes by comparison between signatures the rank of likelihood following the idea that if

two papers use mainly the same concepts they should be similar.

5 APPLICATIONS

We present two applications, one with ISL (International Symposium on Landslides) with a large number of references and bibliometric results, the second with ITA (International Tunneling Association Conferences) with all text papers available providing after processing, a great help for finding similar papers.

5.1 *Contents of the ISL data-base*

We built a database, called base ISL that contains.

- references of the last 6 conference ISL (International Symposium on Landslides) and two other Congress on slope stability. (Lausanne 1988, Christchurch 1992, Trondheim 1996, Cardiff 2000, Rio de Janeiro in 2004, Xi'an, 2008, and also Prague 2002 and Shikoku 1999)
This represents nearly 2000 references of articles.
- proceedings of Xi'an are on CD-ROM, with 270 pdf files available. They increase the data-base by their references of the articles and we can also calculate the "signatures" of these articles.
The references (bibliography) of these articles with full text were automatically red to provide references by keeping the relationship "reference quotes - cited reference," which will identify communities of researchers. (3110 references added)
- other references are set in the data-base coming from Internet researches on editor's web-site, from an excerpt of Geofind, and from large bibliographies from state of the art papers for example. Here we use the references given by S. Leroueil in his Rankine lecture.

Congress	Kind of document	Nb of references from documents	Nb of references set in DB	Nb of authors encountered	Nb of authors set in DB	Nb of expressions encountered
ISL 5 1988	Table of contents	239	239	507	449	
ISL 6 1992	Table of contents	289	285	700	516	
ISL 7 1996	Table of contents	315	315	775	566	
Shikoku 99	Table of contents	227	227	638	395	
ISL 8 2000	Table of contents	276	276	668	440	
Prague 2002	Table of contents	105	105	281	151	
Rankine lecture 2003	Bibliography	268				
ISL 9 2004	Table of contents	234	233	774	499	
ISL 10 2008	CD	270	3110 (1)	10090	4064	216802 (2)
Geofind	Data-base	3136	3116(4)	9515	4450	
Journal (3)	Web sites	301	301			

Table 2 : Some statistics when building the ISL knowledge-base.

(1) This large number shows the supply of all bibliographies at an average of 11 cited references for each paper. (max = 102)

(2) Average number of expressions in signature = 112 (min = 43, max = 378).

(3) Geotechnique, Canadian Geotechnical Journal., Japanese Geotechnical Journal, Soils Dynamics, Computer and geotechnics. We interrogate only the year 2010.

(4) The number of reference set in the data-base depends of the time when the references are presented, because each reference is set only one time. For this number, it was just after setting ISL 10.

5.2 *Some results from the ISL knowledge base*

- The number of identified authors is more than 10000 in the knowledge-base. As each author can be found several times, the number of authors in the data-base is less than the number of authors encountered when processing.
- The most cited papers are from Hungr, Bishop, Duncan and Barton.
- During the eight ISL congresses 10 authors contributed to more than 20 papers. The total of the papers written by the 100 most referenced authors was 1211, and 2554 authors wrote only one paper

5.3 *Contents of the ITA knowledge-base*

We use the ITA conferences of Istanbul (2005), Aggra (2008), Prague(2007) and Budapest (2009). The taxonomy is built with the Seoul conference(2006). 978 papers get a signature that allows an efficient search for similar papers.

In table 3, 12 lines or an output from MKD are printed. The sequences to obtain this result are:

A query with three words selects the papers with these three words in the title and a list of references (papers) is set.

Giving a paper reference, MKD computes ratio, cosine of the two vectors representing each paper (cos), comparing the given reference to each reference of the list.

For the two last columns (log likelihood and Tf-Idf) the comparison of each paper is done with the whole corpus of all the references of the list. (this table is ordered with cos values)

num	id_ref (paper)	signature	nb of common syntagms	ratio %	cos	log likelihood	tf-idf
3	777	111	45	45.96	0.66	58.48	86.11
6	13	73	44	44.94	0.66	35.68	45.71
16	464	70	40	40.86	0.66	37.73	48.73
17	468	93	42	42.9	0.66	51.97	75.55
1	199	92	47	48	0.64	45.67	62.51
10	17	103	48	49.02	0.64	43.39	60.36
11	144	99	38	38.82	0.64	43.58	59.53
15	463	80	38	38.82	0.64	44.61	60.92
18	474	86	38	38.82	0.64	35.07	45.26
19	475	77	39	39.84	0.64	32.9	42.11
0	176	71	31	31.68	0.63	42.31	57.58
4	904	69	33	33.72	0.61	31.21	37.43

Table 3 : Sorting papers of ITA

5.4 *The use of MKD in the validation process*

We have got by e-mail the wishes of binomial who want to work on a particular subject. With the words of the subject we ask the knowledge-base and sort the titles of the papers with these words in them. We can have no answer (words are too accurate and they are not in the title), or too many titles (the words are too common). Then, we use query with all the words comparing to all the signature of the data-base and we obtain for each title a ratio showing if the paper suits the question.

An other way to use MKD, which I prefer, is to give to MKD the name of a paper that, because we have read it, or because we know the author and their favorite subject, meets to the student requirements.

MKD is able to give for all papers in the data-base ratios of likelihood with this paper. We chose the two or three first papers and give to the students the links as they can download the papers.

Reading the texts the binomial can analyze them. The document that the binomial sends back to me, is not only a translation and an abstract, question like: does the method described seems mature? What are the results? is the evaluation fair? all assumptions are they really validated? What seems new and with good perspective? Etc

This new methodology for course credit validation is appreciated by the students, who often are so enthusiastic that they submit work over and above that which is required. Less motivated students at least do the translation of the papers into French, thereby assimilating the concepts contained therein. And for the professor the correction of the different home works is not boring as each work is original and it is quite easy to evaluate the students when their reasoning and references to the syllabus are clearly set on the sheet.

5.5 *Other use*

In order to define an order in the reading of Xi'an papers, we set as list of expressions the signature of one of our papers presented at this conference. Asking MKD with this list, ratio values are from 48% to 79%, except when the signature checks its proper file, where the result is 100%. We know so, an order for reading the 270 papers of the conference, regarding our own interest.

6 LOOKING NEXT FUTURE

This knowledge-base is one step towards the search and storage of knowledge. After research about expert-systems, (Faure R.M. et al, 1992), then about distributed data-base (Faure R.M., 1999) , since 2003 we have undertaken work in this area, first applied to the field of tunnels (Faure R.M. et al, 2007), then open to other areas in the current draft MKD (Management of Knowledge Domains).

Below we indicate, after a reminder in relation to knowledge, some information about the coming modules of the project.

A reminder on brief definition of knowledge

- Knowledge can only be expressed if one uses it (we have to open an encyclopedia to find something), it is highly contextualized (in this case, we must do this).
- The knowledge is expressed in a deductive form (order logic 1) as opposed to descriptive. In one paper much of it is descriptive and few sentences are deductive.

We call text fragments portions of text (one or two sentences) that contain a deductive form that looks like: if, in this context, then these consequences. After an automatic research we have causal relationships between two contexts. They are extracted from the text of the article, and stored in the knowledge-base as a piece of text, and MKD transforms these text fragments in granules of knowledge.

We designed the granule of knowledge (Faure N., 2007) to handle and store knowledge very effectively. The granule of knowledge can be represented by its properties.

- Structure of the granule.

Structure of the granule is ternary (contexts, relationships, and universe) to better express

knowledge, with additional information (origin, signature, etc ...). The two contexts are linked by a causal relation chosen among five that are: information, recommendation, obligation, warning, and negation. The definition of a universe allows the use of dimensions and sizes. (Faure N. et al, 2008)

● Uniqueness of the granule.

A granule is unique. When transforming a text fragment into a granule, its neighbours are found and a new granule is created only if it brings new knowledge. It can also change, during this process, when information completes a pre-existing granule. This eliminates the redundancy of causal relationships, they are stored only once, but the reference to the paper contributing to a granule is retained in order to find articles that deal first with this relationship.

● Composition of granules.

The structure of the granules allows composition and thus the writing of new knowledge, or the return of complete theories. This composition is based on a mapping driven by a measure of semantic proximity, similar to those used for the mappings between ontologies.

– With this new concept, already implemented during the Ramcesh project, (Faure et al., 2007) (Faure et al, 2006), we have a powerful tool that answers at the recurrent question: in that case, what I have to know? MKD, using the granule of knowledge allows reactivity, completeness, and safety checking during design phases.

7 CONCLUSION

The French Committee of Soils Mechanics (CFMS) recently presented a web-site: www.geotech-fr.org, allowing the access to more than 30000 papers (pdf files). In the future the coming papers will be treated with tools like MKD for a better dissemination of knowledge with more accurate purposes. (Mestat, 2011)

Teaching is also a very important challenge and storing and formatting knowledge will be useful for teachers. (Faure, Thimus, 2004)

At an international level (ISSMGE, Innovation and Development Committee) lot of ideas deals also with these topics of a better connection between members of our community. (Geoworld, Webinar, are also examples).

MKD shows some of the opportunities that will enable knowledge management and its use. No doubt, the future will be full of innovations in this area and it is therefore urgent to consider it.

8 REFERENCES

Beney J., 2008, Classification supervisée de documents. *Hermès-Lavoisier,* 181p

Dunning T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics, Volume 19, number 1*, pp. 61-74

Faure N. (2007). Un système d'aide à la modélisation des connaissances en géotechnique. *Thèse, Université Lyon 3*

Faure N., Faure R.M., Balasch M.A., Cottaz Y., 2008, Mise en forme de granules de connaissances à l'aide de l'outil RAMCESH, *congrès International AFTES, Monaco.*

Faure R.M., Mascarelli D., Vaunat J., Leroueil S., Tavenas F.*;* 1992, Present state of development of XPENT, expert system for slopes stability problems.; *6th Int. Congress on Landslides Bell editor Christchurch.* pp1671-1678

Faure R-M., 1999, Data-bases and the management of landslides. *Int. Symp. on Landslides. Shikoku (Japan). IS Shikoku'99*, pp1317-1330

Faure R.M. & Thimus J.F. 2004. Contribution of on line tools on Internet for the teaching of slopes and tunnels stability, EurEnGeo 2004, *First European IAEG Conference*: 59-69. Liège.

Faure R.M., Faure N., Hémond G. 2006. The use of knowledge management in the management of tunnels and tunnel projects*. ITA-AITES World Tunnel Congress.* Séoul

Faure R.M., Faure N., Hémond G. 2007, Recueil Assisté et Maniement des Connaissances des Espaces Souterrains Habités. *Tunnels et Ouvrages Souterrains*, n°202

Mestat Ph., 2011, Ouverture du site internet « Géotechnique Francophone », *Lettre de la Géotechnique n° 57.*