Usage des 'big-data' pour la validation d'un cours de Mécanique des sols.

The use of big-data for course validation

René-Michel FAURE (1)

Professeur, ENTPE, Vaulx en Velin, France

RESUME - La validation d'un cours de spécialisation de mécanique des sols (dernière année d'école d'ingénieurs) est un challenge pour le professeur qui doit permettre à ses élèves de montrer ce qu'ils ont retenu de son enseignement et d'exercer leur jugement sur des cas réels. La synthèse d'articles (en général 3 ou 4) portant sur des sujets voisins devient possible en utilisant une application dédiée fondée sur des techniques de big-data.

ABSTRACT – Course validation is a challenge for specialization courses. An ad hoc code using big-data techniques can help efficiently the teacher in the search for papers dealing with the same topics. Each team of students have to summarise, to analyse and to provide reflexions about these papers. The result is a very good acceptance of this evaluation.

1 I**ntroduction**

Depuis quelques années les big-data sont devenues un incontournable des approches techniques et scientifiques. Comme de nombreuses nouveautés ce vocable recouvre bien des disparités. L'approche présentée dans cet article correspond à la recherche d'une solution à un problème d'enseignement et tout particulièrement celui de la validation d'un cours de spécialisation.

1.1 Problématique

Chargés du cours de spécialisation de Mécanique des Sols en dernière année d'école d'ingénieurs, un des challenges est de prolonger un cours au nombre d'heures insuffisant, le champ des spécialisations en mécanique des sols étant particulièrement vaste. Un autre challenge est d'accroître l'intérêt des élèves pour ce cours en présentant les innovations permanentes de nos techniques, et aussi de proposer une validation rigoureuse qui n'outrepasse pas les capacités de correction du professeur. Ces challenges furent au départ de notre quête pour mettre au point la procédure nécessaire. La tâche est d'autant plus délicate que l'air du temps veut que le savoir soit annexe, il est sur internet, et que l'enseignement se complaît dans le comportemental. L'ingénieur est pour nombre d'établissements, un meneur d'hommes ou un débatteur de projets (qu'il ne maîtrise pas souvent).

1.2 L'idée d'une solution

L'idée est donc de mettre les futurs ingénieurs au pied du mur, face non pas à un projet, mais à plusieurs vues d'un même problème, à travers projets et comptes-rendus de recherche, et de leur demander une réflexion et des commentaires sur des documents qu'il faut leur proposer.

La difficulté de cette approche est donc dans le choix de ces documents variés et portant différents aspects d'une problématique. Pour quelques binômes c'est assez facile, mais pour toute une promotion cela devient impossible. C'est pourquoi, sans le savoir au départ, nous nous sommes engagés dans le maniement des big-data.

La notion de « big » est très relative, mais cela veut dire beaucoup, en tout cas trop pour la mémoire d'un professeur et de ses moyens. Le mot « data » est aussi pluriel car il recouvre toute sortes de données, des valeurs mesurées au comportement de mammifères, sans oublier les aspects temporels.

1.3 Restrictions

Dans le cas présent, on se limite aux données contenues dans les écrits de la profession. Ces écrits sont très nombreux et personne n'est capable de lire toutes les publications accessibles, des différentes rencontres ou congrès. Le système d'aide recherché doit donc être capable de comparer deux articles parmi un large choix, d'évaluer leur ressemblance et de les proposer dans un corpus de textes illustrant une problématique donnée. Le choix de la problématique est laissé aux binômes d'élèves qui à l'aide de mots-clés définissent cette problématique.

Avec un choix bien équilibré des articles, les élèves découvrent combien la modélisation géotechnique est parfois délicate, que les calculs ont des limites et que la technologie va sans cesse de l'avant pour réaliser des projets de plus en plus contraints par l'environnement.

2. Rappel de l'existant

2.1 La déception des moteurs de recherche.

La quête étant bien définie, peut-on trouver facilement et rapidement le petit corpus d'articles correspondant au mieux à une problématique définie par un binôme ?

Les milliards de pages qu'indexent les moteurs de recherche sont visiblement un trop vaste champ de recherche et les résultats sont souvent décevants. Il faut de nombreuses fois ré-exprimer la question pour obtenir un article, puis deux et alors sommes nous certains de la complémentarité de ces articles pour le but recherché ? La réponse est généralement non, les moteurs de recherche ayant des finalités différentes des nôtres, et surtout travaillent avec comme système de classification, la popularité de l'article.

Un autre écueil de cette approche est la propriété des documents et bien souvent il faut payer l'information (même pour voir) ou être affilié (en payant) à un organisme qui revend le travail de chercheurs dépossédés de leurs droits. Nuançons ces propos en notant des initiatives de sociétés savantes, comme le CFMS, qui mettent à disposition des fonds documentaires importants.

Récemment de nombreuses critiques environnementales ont été portées contre les grands datacenters qui usent beaucoup d'énergie à indexer ces milliards de pages, en

attendant les ordinateurs quantiques pour trier encore plus dans des masses de données encore plus grandes. En 2015, les datacenters qui mettent en forme les données de toute sorte consommaient 10% de la consommation électrique mondiale. (rapport Villani)

2.2 La faillite des systèmes experts.

Dans les années 90 la mode et les espoirs portaient les systèmes experts vers la conquête du monde. Si de beaux résultats ont été obtenus dans des domaines fermés (industrie), leur faillite a été totale dans des domaines ouverts (géologie, mécanique des sols) pour lesquels la connaissance n'est pas bornée.

3 Approche utilisée.

Inutile donc d'accéder aux serveurs où sont entassées les données du monde, il est plus efficace de créer un îlot de connaissance fondé sur le contenu des nombreux congrès de Mécanique des sols, facilement récupérables et qui donne la certitude de rester dans notre domaine.

Bien qu'évoquée plus haut la notion de big-data, le projet a donc lieu sur un ensemble de données relativement restreint, dont l'intégralité peut être traitée dans des délais raisonnables (quelques secondes) sur une machine standard. Pour autant, la solution proposée peut être employée sur des corpus beaucoup plus larges sans modification notable. Un fond de plusieurs dizaines de milliers d'articles est tout à fait envisageable.

3.1 Constitution de la base de données et outils utilisés.

Pour une maîtrise complète de l'outil, ce dernier a été construit en utilisant des briques informatiques abordables suivant la démarche du prototypage incrémental où l'outil est en permanence adapté à la demande jusqu'à que cette dernière soit satisfaite. Le langage de programmation est Php associé à une base de données MySql.

Cette base de données a été alimentée par les congrès choisis, ici ceux des JNGG, En pratique un congrès est composé de 200 à 800 articles au format pdf.

Les fichiers .pdf, sont renommés puis transformés en fichiers texte (.txt) pour être traités par un premier outil qui scinde le texte en trois parties, le titre, le texte et la bibliographie. Les titres sont édités pour être validés puis stockés dans la base de données avec le lien vers le fichier .pdf d'origine.

3.2 Nécessité d'un lexique et sa construction.

Partant de l'hypothèse forte que deux articles utilisant les mêmes mots sont semblables, la recherche se porte sur cette ressemblance entre articles. Quelques difficultés surgissent. Les mots ne suffisent pas, il faut aussi traiter des groupes de mots que l'on appelle syntagmes. Par exemple le « degré de consolidation d'un sol » est significatif à l'ingénieur, bien plus que les mots pris isolément. Il y a aussi les synonymes et le renard hydraulique n'est pas un animal mouillé. Nous verrons plus loin comment donner le sens juste à un synonyme.

Il faut donc bâtir un lexique de mots et de syntagmes. Pour cela un outil, libre de droit pour un usage personnel, mis au point à l'université de Stuttgart va retrouver dans un texte d'abord lemmatisé (les mots sont reconnus au singulier), mots et syntagmes suivant des patrons que nous avons choisis. Et ainsi pour chaque texte d'article nous obtenons la liste

des syntagmes qui le composent avec leurs fréquences. Avec la liste de tous les syntagmes d'un congrès, classée par ordre de fréquence, nous obtenons par tri manuel un lexique 'géotechnique' et un anti-lexique. Chaque syntagme est stocké avec la date de l'article, pour plus tard permettre des recherches d'antériorité. Pour le premier congrès traité ce travail portant sur des milliers de syntagmes est un peu fastidieux, mais par la suite seuls les nouveaux syntagmes seront traités.

A partir du répertoire des fichiers lisibles (au format .txt) un module effectue une analyse morpho-syntaxique pour retrouver dans chaque fichier, les concepts du texte. Ces concepts sont représentés par des mots ou des ensembles de mots, jusqu'à 5 mots, qui sont détectés à l'aide de patrons. L'analyse morpho-syntaxique détermine d'abord le rôle syntaxique de chacun des mots du texte, c'est-à-dire si ce mot est un nom (N), une préposition (P), un adjectif (A), un verbe, etc…Puis dans une seconde lecture, l'outil recherche les groupes de mots définis par les patrons. Le choix des patrons nominaux suivants a été fait, à savoir N, NA, AN, NPN, NPAN, NPNA, NPNPN et Nom Propre. Pour chaque groupe de mots correspondant à un patron, l'analyse compte le nombre de fois que ce groupe de mots, appelé syntagme, apparait dans l'article.[1]

Ainsi, pour chaque article, une liste de syntagmes est créée, avec attaché à chaque syntagme son occurrence (le nombre de fois), son patron ainsi que la langue et l'année qui ont été attribuées de façon automatique à l'article.

Avec ces syntagmes un lexique est alors établi. Pour chaque congrès on obtient un lexique dont le nombre de syntagmes est assez important. L'outil permet de comparer ces lexiques et surtout de les additionner dans un lexique plus vaste associé à un ensemble de congrès. Par exemple, pour les six congrès JNGG (2002 à 2012) le nombre de syntagmes détectés approche 150 000. La construction des lexiques de chaque congrès permet d'attacher à chaque syntagme une donnée importante : le nombre d'articles dans lesquels il est utilisé. Cette donnée va permettre, combinée avec l'occurrence de chaque syntagme, d'ordonner par « usage décroissant » tous les syntagmes du lexique. Cette probabilité d'usage s'appelle le tf-idf [2] et par simple seuillage on retient dans le lexique attaché au domaine couvert les 'n' syntagmes les plus employés[3]. Ce nombre 'n' est fourni par l'utilisateur. Plus 'n' est grand, plus les comparaisons seront précises. L'usage montre qu'un lexique d'environ 2000 syntagmes est bien adapté aux buts poursuivis.

La détermination d'un lexique va permettre l'indexation automatique de chaque texte, c'est-à-dire lui affecter un certain nombre de « mots-clés » qui sont les syntagmes communs à l'article et au lexique. Nous appellerons cette liste de syntagme, la signature de l'article.

3 3 La signature d'un article.

[1] Exemple de patrons morpho-syntaxique : N : Nom : glissement, NA : Nom + Adjectif : glissement lent, AN : Adjectif + Nom : grand glissement, NPN : Nom + Préposition + Nom : glissement de terrain, NPAN : Nom + Préposition + Adjectif + Nom : glissement de grande ampleur, NPNA : Nom + Préposition + Nom + Adjectif : glissement de terrain rocheux, NPNPN : Nom + Préposition + Nom + Préposition + Nom : glissement de terrain dans l'argile

[2] C'est une fonction simple du nombre d'occurrence du syntagme (combien de fois ce syntagme est utilisé dans tous ces articles) et du nombre d'articles dans lequel est utilisé ce syntagme.

[3] La fonction "log-vraisemblance" est aussi utilisée (loglikelihood measure).

C'est la liste des syntagmes présents dans l'article et dans le lexique. Pour des articles de 6 à 10 pages cette signature comporte de 200 à 800 syntagmes. Il est alors facile de comparer sémantiquement plusieurs articles en recherchant les proportions de syntagmes communs entre deux signatures. Un usage annexe est par exemple de comparer un article (le sien) avec tous les articles d'un congrès, (ou avec tous les articles en base) de les classer par ordre de ressemblance (de l'ordre de 3 secondes avec un Pentium5 pour 3000 articles) et de ne lire alors que les premiers du classement, pour connaître, ainsi qui travaille sur le sujet.

3.4 Intelligence et algorithmes.

Au vu du résultat la machine peut apparaître intelligente, mais elle n'exécute qu'un algorithme de tri, celui décrit ci-dessus. Il est exécuté avec des temps de calcul très raisonnables car l'espace de recherche est délimité. Mieux vaut chercher une aiguille dans une botte de foin que dans un grand fenil.

Du fait de l'abondance des textes, donc de nombreuses redondances, l'approche classique des documentalistes est en échec, car le système à base de quelques mots clés est insuffisant et la trop longue liste des documents sélectionnés est inexploitable. Avec les signatures tout se passe comme si la recherche se faisait avec des centaines de mots-clés générés et utilisés automatiquement.

L'approche présentée ici est ainsi située entre la recherche plein texte et la recherche documentaire grâce au lexique du domaine qui cible les bons concepts.

4 Quelques résultats.

Le traitement des différents congrès JNGG, conduit aux tableaux suivants qui indiquent le nombre des éléments traités. La base comporte 4123 références d'articles et 7092 auteurs identifiés. (tableau 1). Le traitement a trouvé 149163 syntagmes, y compris les doublons, dans les articles ce qui permet la détermination du lexique avec un simple tri car les 4454 syntagmes différents sont ordonnés suivant leurs tf-idf et seuls les 1762 premiers sont retenus (tableau 2).

L'évolution du vocabulaire est sensible montrant que le domaine couvert par les journées s'élargit.

Tableau 1 : Analyse de six congrès JNGG

Congrès	Fichiers initiaux txt : Nombre de lignes			Fichiers S1			Fichiers S2				Fichiers S3
	S1	S2	S3	Nb articles	Nb auteurs lus	Nb auteurs nouveaux	Nb ref lues	Nb ref nouvelles	Nb auteurs lus	Nb auteurs nouveaux	Nb syntagmes lus
Nancy 2002	675	2670	19445	87	184	158	584	570	1021	829	19444
Lille 2004	669	2081	18814	59	185	160	448	420	1034	732	18814
Lyon 2006	899	2813	25621	82	244	202	602	566	1032	927	25620
Nantes 2008	1099	3006	20739	97	307	196	631	601	1372	1008	20739
Grenoble 2010	1503	4541	34699	124	449	247	961	868	2364	1455	34698
Bordeaux 2012	1434	3178	29848	112	423	244	655	557	1694	971	29848
Totaux											149163

Tableau 2 : recherche des lexiques des six congrès JNGG

Congrès	Nb syntagmes lus	Nb de syntagmes retenus	Nb syntagmes sommés	Nb de syntagmes ajoutés
Nancy 2002	19444	1695	1695	1695
Lille 2004	18814	1394	2295	600
Lyon 2006	25620	1575	2862	587
Nantes 2008	20739	1653	3405	522
Grenoble 2010	34698	1848	3969	564
Bordeaux 2012	29848	1750	4454	485

Le domaine des tunnels a aussi été abordé, (Faure et al, 2014)

4.1 La recherche bibliographique.

Elle commence par des tris sur les mots du titre et des noms des auteurs, puis elle peut être poursuivie en utilisant des sous-domaines représentés par des listes de syntagmes. Elle est fondée essentiellement sur des recherches de ressemblance entre article ou même la recherche d'articles semblables dans un ensemble d'articles.

4.2 Extension de l'approche lexicale

Un lexique est construit à partir de l'analyse lexicale de chaque article, pour chacun des congrès. Ces lexiques sont rassemblés dans un lexique général et le tf-idf de chacun des syntagmes est calculé, ce qui permet d'ordonner ceux-ci, et le lexique retenu correspond aux premiers syntagmes de la liste (ici 1762)

Assisté par la machine, on peut repartir ces syntagmes dans des classes, ce qui correspond à des sous-domaines. [4] Voir liste des sous-domaines [5]

Il est alors facile de rattacher chaque article à un sous-domaine par simple comparaison de syntagmes. On obtient ainsi une thématisation d'un ensemble d'articles. Des opérations d'addition, de soustraction entre listes, permettent, sans écrire un seul syntagme, d'obtenir un ensemble de syntagmes représentant une idée, un concept, une théorie, un ouvrage très particulier, etc…

Le classement par degré de ressemblance permet d'identifier les articles qui se ressemblent le plus.

Le tableau 3 liste les syntagmes du sous-thème pente pour accéder à une thématisation de l'ensemble des articles.

Tableau 3 : Liste des syntagmes d'un sous-thème

9	pente	19 25	_friche_glissement_mouvement_pendage_pente_protection_redan_remblai_remblaiement_remblayage_remodelage_replat_reptation_risberme_rive_stabilisation_talus_tranche_versant_courbe de rupture_glissement de terrain_grand pente_hauteur de chute_hauteur de remblai_ligne de rupture_mouvement de sol_mouvement de terrain_mouvement de versant_ouvrage de protection_ouvrage en sol_pente de talus_pied de talus_pied de versant_plan de cisaillement_plan de fracture_plan de glissement_profil de versage_sol de remblai_surface de glissement_surface de rupture_temps de stabilisation_type de rupture_zone de transition_zone en mouvement

Une analyse temporelle, c'est à dire la recherche d'antériorité d'un concept est possible, à chaque syntagme étant attaché la date la plus ancienne quand il apparait dans l'analyse des textes.

[4] Une classe particulière est nécessaire, celle qui correspond aux syntagmes de sens 'général', sans rapport avec un sous-domaine.

[5] Liste des sous-domaines (liste non limitative) : Géologie, Hydrogéologie, Mécanique des sols, Risque, Matériel, Travaux, Tunnels, Fondations, Pentes, Propriétés, Phénomène, Mécanique des roches, etc…

La synonymie est facilement résolue avec cette partition en sous-domaine. En cas de doute, tester les syntagmes voisins permet de connaître le sous domaine, le renard de mécanique des sols ne peut être celui de la faune. L'articulation des syntagmes dans un réseau de sous-domaines permet d'esquisser une sémantique d'ensemble, suffisante pour autoriser une recherche robuste sans impliquer de biais de représentation. (Faure et al, 2017)

5 Solution au problème posé et résultats.

Pour les élèves, à partir de la série de mots-clés qu'ils fournissent, (ils ont accès au lexique) la recherche du professeur est immédiate, un simple contrôle lui suffit pour valider le travail demandé (synthèse, commentaires et réflexions) portant sur les articles soumis.
Tout cela se fait à travers les réseaux et permet gain de temps et discussion, quand un élève saisit mal le contenu d'un article.
Les résultats sont très encourageants. Pour de nombreux élèves, c'est la découverte que la mécanique des sols n'est pas confinée dans un laboratoire, qu'il y a bien d'autres équations que celles du cours et que l'innovation sur le terrain est un très puissant moteur de réflexion. Cela les amène bien souvent à dépasser leur problématique initiale en la situant par rapport au champ disciplinaire dans son ensemble et quelques-uns s'engagent alors « en mécanique des sols ».

5.1 Pour aller plus loin, connaissances et compétence.

Créer la compétence c'est, sur un problème, apporter tout ce que l'on doit savoir, c'est à dire un outil qui indique les bonnes pages du savoir, les bonnes pages d'une encyclopédie.
Pour cela on se rappelle l'approche système expert avec ses règles de production. Une règle de production est une phrase en trois parties. Les prémisses qui explicitent le contexte initial, une charnière de type déductif, et une conclusion qui décrit un nouveau contexte, soit en résumé, « **si** dans ce cas, **alors** conclusion ». La faillite des systèmes experts venait de la recherche de ces règles, (ou mise en forme du discours de l'expert qui veut bien se plier à cet exercice) qui, jamais complète, ne pouvait aboutir.
Dans notre environnement, (grâce à un lexique partitionné) et quelques outils de traitement automatique du langage, la recherche automatique et l'analyse de phrases déductives permet d'écrire la connaissance de façon formelle et maniable. C'est le granule de connaissance. (Faure et al,2007, 2008) L'outil peut alors répondre à la question : dans ce cas que dois-je savoir ? La population de granules croissant avec l'analyse de nouveaux textes, les réponses seront de plus en plus complètes, évitant oublis et erreurs, fournissant à l'utilisateur une aura de compétence.

6 Conclusion

La réalisation informatique de cette approche conduit à quelques réflexions.
Si la collecte automatique de données (surveillance sismique, forages, etc.) conduit à des big-data nécessitant des outils surpuissants (deep-learning), il ne faut pas oublier d'exploiter la connaissance textuelle, *qui est enrichie de la réflexion d'un auteur* par rapport à des données brutes.

Pour un thésard qui nourrit la machine de nombreux congrès ou paquets d'articles, le traitement fait par la machine lui procure un gain de temps important et lui assure de ne pas oublier le moindre détail.
L'usage de l'outil, dans le cadre d'un cours est, comme écrit plus haut, tout simplement extraordinaire, car elle délivre le professeur d'un travail fastidieux et permet à l'élève de s'approprier un travail qui lui fait profondément découvrir l'intérêt et l'étendue la Mécanique des Sols.

7 Bibliographie

Faure N. 2007, Un système d'aide à la modélisation des connaissances en géotechnique. Thèse, Université Lyon 3, pp1-151.

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 Int. AFTES, Monaco, pp. 505-514.

Faure N., Thimus J.F., Faure R.M., 2014, Analyse statistique lexicale pour la construction d'une base de connaissances, Tunnels et Espaces Souterrains, n°244, pp317-328.

Faure N., Faure R.M., 2017, Représentation informelle de connaissances issues de documentation technique : une expérience. I2D, Information, données et documents n°1, pp70-79.

Site: www.pentes-tunnels.eu/mkd