Problèmes de l'analyse automatique de textes

Autor(en): **Margot, J.-M.**

Objekttyp: **Article**

Zeitschrift: **Nachrichten / Vereinigung Schweizerischer Bibliothekare, Schweizerische Vereinigung für Dokumentation = Nouvelles / Association des Bibliothécaires Suisses, Association Suisse de Documentation**

Band (Jahr): **48 (1972)**

Heft 2

PDF erstellt am: **21.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-770960

allein nicht mehr gelöst werden. Arbeitsteilung und Gruppenarbeit sind unabdingbare Voraussetzungen für das Gelingen derartiger Vorhaben. Die Dokumentalisten können nicht einfach Forderungen aufstellen und die Lösungen den EDV- bzw. Mikro-Film-Fachleuten überlassen. Man muß von Anfang an zusammenarbeiten und fortwährend miteinander sprechen. Eine Ausbildung der Dokumentalisten in Richtung EDV und vermehrtes Verständnis der Informatiker für die Bedürfnisse der Dokumentation werden unumgänglich. Ein Erfahrungsaustausch im Sinn des heutigen Seminars liefert daher eine willkommene Gelegenheit, dieses gegenseitige Verständnis zu fördern.

PROBLEMES DE L'ANALYSE AUTOMATIQUE DE TEXTES

par *J.-M. Margot,* IBM, Berne

Introduction

Dans un système documentaire, manuel ou mécanisé, un document (livre, article de périodique, brevet, carte ou plan, rapport, etc.) est toujours représenté par deux éléments. La description physique du document correspond à la notice catalographique où figurent les coordonnées (titre, auteur, éditeur, date, nombre de pages, etc.) permettant de retrouver de manière univoque le document ainsi signalé. La deuxième partie sert à permettre la recherche documentaire. Il s'agit de l'image du contenu du document, un ou plusieurs codes de classification (Classification Décimale Universelle par exemple), des mots-clés, un résumé ou même un texte complet du document si celui-ci n'est pas trop long. Les codes de classification et les mots-clés ont fait l'objet des deux précédentes conférences et mon propos est de présenter quelques problèmes à résoudre si l'on désire utiliser un ordinateur pour analyser et traiter des résumés ou le texte complet d'un document. L'analyse automatique de texte a pour but de:

– sélectionner certains éléments d'un texte sans les changer,
– transformer la forme ou le contenu du texte sans en modifier le sens.

Des exemples vont mettre en évidence quelques-uns des problèmes susceptibles de surgir lors de l'analyse automatique de textes.

Les problèmes présentés peuvent être résolus et l'ont été, mais nous n'avons pas l'intention, dans le cadre de ce modeste exposé, de montrer les solutions existantes, car cela exigerait plus de temps et ferait appel à des notions d'informatique spécialisée.

Analyse avec transformation du texte

Le texte lu par l'ordinateur est traduit en un nouveau texte, rédigé en langage documentaire, à la suite d'une analyse grammaticale automatique. Il en résulte un nouveau document qui, selon le degré de perfectionnement de l'analyse, peut être présenté sous forme de graphes, de mots-clés ou de résumé. De telles analyses en sont encore au stade du développement et de la recherche et ne connaissent pas encore une utilisation répandue. Extraire d'un texte les quelques phrases les plus significatives pour en représenter un résumé ne pose pas des problèmes insolubles, mais par contre la rédaction d'un résumé à partir d'un texte signifie une analyse sémantique du contenu et un tel procédé reste par conséquent très aléatoire avec un ordinateur.

Ce qui est déjà plus à portée de l'informatique actuelle, c'est de représenter un document par des mots-clés ne figurant pas dans le texte.

Par exemple, le texte suivant: «*Ein Generalstreik hat am Mittwoch Venedig und seine Umgebung lahmgelegt*» peut être représenté par les mots-clés «*Generalstreik*», «*Venedig*», figurant déjà dans le texte, mais aussi par les mots-clés «*Streik*», «*Italien*», «*Europa*», «*Staedte*», «*Soziale Konflikte*», «*EWG*», qui n'y figurent pas. Faire indexer un texte par un documentaliste ne pose que des problèmes de place de travail, de personnel, de salaire, etc. Par contre, faire indexer le même texte par un passage de celui-ci en ordinateur pose des problèmes d'analyse et d'organisation, surtout si le texte contient beaucoup de mots composés, de signes de ponctuation, et s'il est rédigé en plusieurs langues. Quelques-uns de ces cas sont examinés plus loin avec plus de détails.

Analyse sans transformation du texte

Comme nous venons de le mentionner ci-dessus, la méthode la plus simple pour un ordinateur pour extraire les mots significatifs d'un texte consiste à comparer une liste de mots avec chaque mot du texte. Mais déjà un problème se pose: qu'est-ce qu'un mot dans un texte? La première réponse qui se présente à l'esprit est la suivante: un mot est séparé du précédent et du suivant par un caractère blanc. Mais à peine a-t-on formulé cette affirmation qu'on s'aperçoit que l'expression «*L'Incident*» contient deux mots non séparés par un blanc. Alors on peut compléter la réponse ci-dessus en affirmant que cette séparation est réalisée par un blanc ou un signe de ponctuation. Mais en poussant la réflexion un peu plus loin, on s'aperçoit que «*Valery Giscard d'Estaing*» et «*La Chaux-de-Fonds*» sont deux mots, deux descripteurs chacun composé, si on respecte la règle ci-dessus, de quatre mots en realité. Ces quelques réflexions montrent que l'analyse automatique de textes ne peut être résolue facilement, même si, au premier abord, la technique à utiliser paraît

simple et immédiate, comme dans l'exemple suivant: «*La danse des abeilles, très riche en information, présente un intérêt exceptionnel. C'est grâce à cette danse que les ouvrières qui ont trouvé une nouvelle source de nourriture font savoir dans quelle direction et à quelle distance elle se trouve*».

Un programme d'ordinateur peut comparer ce texte avec une liste de mots sans signification *(a, afin, alors, c, car, ce, cette, dans, des, donc, elle, elles, en, est, et, il, ils, la, le, les, mais, ni, ont, ou, par, pour, que, qui, sans, se, sont, très, un)* ou avec une liste de mots-clés *(abeille, abeilles, apiculteur, apiculture, fleur, fleurs, forêt, forêts, hausse, hauses, mâle, mâles, miel, ouvrière, ovrières, rayon, rayons, reine, reines, ruche, rucher).*

Le résultat du premier traitement fournira la liste des mots-clés suivants: *abeille, danse, direction, distance, exceptionnel, font, grâce, information, intérêt, nourriture, nouvelle, ouvrières, présente, quelle, riche, savoir, source, trouve.*

Le résultat du second traitement fournira les mots-clés suivants: *abeilles, ouvrières.*

Cet exemple a été choisi de manière à montrer quelques-uns des problèmes à résoudre lorsqu'on désire entreprendre une extraction automatique de mots-clés dans un texte:

– le choix du procédé (liste de mots-stop ou de mots-clés),
– le choix des mots dans la liste de référence,
– les différentes formes grammaticales d'un même mot.

C'est ainsi que le mot «*grâce*» sera un mot-clé dans un lexique théologique; mais un mot-stop dans le texte ci-dessus, dont l'analyse aurait dû fournir par exemple le résultat possible suivant: *abeilles, communication, danse, distance, nourriture, orientation, ouvrières.*

Par l'étude et la réflexion personnelle au sujet d'un tel exemple simple, on constate que l'analyse automatique de textes exige la solution sous forme d'algorithmes de bien des problèmes, afin de pouvoir être utilisée de manière rentable et de manière à répondre aux besoins des ingénieurs et des chercheurs.

Les problèmes de la saisie des données

Plus un texte est long, plus il faut de temps pour l'écrire. Cela est vrai pour une dactylo aussi bien que pour une perforeuse. Dans le cas d'un traitement automatique de texte, on ne peut négliger cet élément. Les textes fournis à l'ordinateur doivent pouvoir être lus par lui et cela implique le choix du support matériel du texte, bande magnétique, bande de papier perforé, lecture optique, cartes perforées ou ligne téléphonique dans le cas de l'entrée des

documents par terminal (machine à écrire ou écran de visualisation connecté directement à l'ordinateur). Evidemment que la solution la meilleure serait que tous les documents édités et circulant à la surface du globe existent en deux versions, l'une lisible par l'homme (livres, revues, microfiches, etc.) et l'autre lisible par la machine. Mais nous sommes encore loin de cet état de faits où les éditeurs produiraient simultanément chaque document sous deux formes. Cependant des réalisations dans ce sens existent déjà; il s'agit des abonnements aux revues signalétiques sur bande magnétique, comme *Chemical abstracts, Engineering index,* etc.

Nous n'en sommes pas encore à disposer de livres, de périodiques ou de brevets sur bande magnétique. Chaque centre de documentation doit donc se préoccuper de l'écriture de documents sous forme lisible par l'ordinateur. Quelques chiffres permettent de se faire une idée de l'investissement nécessaire: Une bonne dactylo frappe 2000 à 4000 caractères à l'heure, une perforeuse 4000 à 6000. En tenant compte de la fatigue, des pauses, des erreurs à corriger, etc., on peut admettre pour une journée de travail de 6 heures un chiffre moyen de 10 000 frappes par jour. Si un document contient 250 caractères, une personne peut écrire 40 documents par jour ou 800 documents par mois.

De tels calculs permettent d'évaluer et de prévoir les coûts et le personnel nécessaire à l'entrée des documents en ordinateur.

Les problèmes dûs à la langue

Le premier problème est celui de la reconnaissance automatique de la langue dans laquelle est écrit un texte ou une phrase. En d'autres termes, comment savoir que «*L'électronique a connu depuis 1945 une croissance fulgurante*» est écrit en français ou que «*Der Berner Lehrerverein hat sich gegen den Vorschlag der Grossratskommission ausgesprochen*» est rédigé en allemand? Bien sûr que cela est facilement résolu si on fait précéder chaque phrase d'un code représentant la langue. Mais cela n'est pas toujours possible, en particulier lorsque des citations en langue originale figurent dans le texte à analyser.

Il convient donc de distinguer les textes écrits dans une seule langue des textes multilingues.

Dans le premier cas, le problème de la synonymie entre mots et même entre phrases se pose très rapidement. Par exemple «*Laut Angaben eines Militaersprechers in Tel-Aviv . . .*» et «*Ein Sprecher des Verteidigungsministeriums in Tel-Aviv hat mitgeteilt . . .*» sont synonymes, mais les mots employés sont différents dans les deux cas. De plus, le problème des différentes formes grammaticales se pose aussi lors de l'indexation automatique de textes. Par exemple, «*Besprechen*», «*Besprechung*» et «*haben besprochen*» sont trois

formes de la mêmes notion, mais chacun de ces termes possède la racine «*Bespr*». Cela permet lors du questionnement de rechercher uniquement sur la racine du mot. Par exemple, un économiste désire connaître tout ce qui a paru en Suisse entre 1965 et 1971 en fait de statistiques de la production et des exportations de réveils électriques. Sa question posée à l'ordinateur pourra avoir l'allure suivante: *(Réveil* 1) (Syn) avec (Electri* ou elektri*) et statisti* et Schweiz* (Syn) et date IL 1965, 1971*. Comme *Réveil* peut apparaître dans un texte au singulier ou au pluriel, il est nécessaire de masquer la terminaison avec le signe* mais ce masque ne peut avoir qu'une longueur de un caractère afin d'éviter que les textes contenant *réveiller* ou *réveillon* soient choisis pour figurer dans la réponse. Comme le mot-clé *réveil* est français et que l'on tient à obtenir aussi les textes allemands et anglais, il faut demander au système de tenir compte des synonymes à l'aide de (Syn). Le connecteur *avec* signifie que le mot *électri** ou *elektri** doit figurer dans la même phrase du texte que le mot *réveil* 1* pour que le document puisse figurer dans la liste bibliographique fournie par le système en réponse à la question. Comme la racine *électri* est valable pour l'anglais et le français, on ajoute la racine allemande *elektri* reliée à la première par la connecteur *ou*. Le troisième mot-clé de la question est *statisti** et comme cette racine est valable pour les trois langues, on ne tient pas compte des synonymes et la terminaison masquée n'est pas limitée en longueur. Les mots-clés qui précèdent et qui suivent les connecteurs logiques *et* doivent être présents dans le texte afin que le document figure dans la réponse. A l'aide de *Schweiz* (Syn)*, tous les documents qui contiennent *Helvetie, Helvetia, helvetique, helvetiques, helvetisch, helvetische, Schweiz, schweizerische, schweizerischem, Suisse, suisses, Swiss, Switzerland,* etc. seront sélectionnés par le système pour figurer dans la réponse à la condition bien entendu, que dans leur texte figurent aussi les autres éléments imposés par la question. Le dernier élément de la réponse signifie que la date du document (pour être sélectionné) doit être à l'intérieur des limites (IL) 1965 et 1971.

C'est ainsi qu'avec une écriture très simple et un effort minime, des questions très précises et assez complexes peuvent être posées à un ordinateur capable d'y répondre dans un temps très court, de l'ordre de la seconde ou de la minute.

Dans le cas de textes écrits en plusieurs langues, le problème des homographes devient aigu. Par exemple, «*reine*», «*quelle*», «*Seine*», «*vers*», «*ton*» sont des mots aussi bien français qu'allemands, «*case*», «*the*» sont anglais et français et enfin «*war*» est anglais et allemand. Une autre difficulté surgit lorsque plusieurs langues ont servi à rédiger un texte. Par exemple, «*Er benutzte anlaesslich seines offiziellen Besuches die Gelegenheit, de rompre une lance en faveur de la conférence de sécurité européenne*» est une phrase écrite en deux langues.

Les problèmes de la lecture elle-même

Plus haut, il a déjà été question de la lecture d'un texte par un ordinateur. Certaines règles de lecture doivent être fournies à la machine afin qu'elle puisse distinguer un mot d'un autre, une expression d'une autre, la fin d'une phrase, etc. Quelques exemples vont montrer la difficulté de définir des règles valables de manière générale, sans dépendre du genre de texte ou de la langue dans laquelle il est rédigé. La différence entre mots-clés ou mots significatifs et mots-stop ou mots non porteurs de sens a déjà été mentionnée. Cependant, dans le texte «*Quant aux fruits, ma femme, qui fait son marché avec un petit carnet . . .*» le mot *marché* fait partie de l'expression *faire son marché* qui peut être considérée soit comme expression-clé ou expression-stop selon les besoins. Par contre le mot *marche,* lorsqu'il figure dans l'expression *marché économique* ou *étude de marché,* est alors à coup sûr un mot-clé. Une analyse précise du texte est nécessaire afin d'éviter des déboires au chercheur qui a posé une question claire et précise et qui reçoit en réponse une liste de documents dont la plus grande partie ne le concernent pas. Les problèmes de ponctuation peuvent provoquer des erreurs encore plus graves. Par exemple, le point est généralement utilisé pour désigner une fin de phrase. Mais il sert aussi à limiter une abréviation ou un jour dans une date en allemand, par exemple «*La R.A.T.P. et la S.N.C.F. ont rajuste leurs tarifs*», «*Der Erzbischof von Granada, Mons. Emilio Benavent . . .*», «*Am 3. Juli in Bern hat . . .*». Le trait d'union n'est pas utilisé seulement pour séparer un mot en syllabes à la fin d'une ligne de texte, mais pour remplacer un mot-clé entier, comme par exemple dans «*Eine Bearbeitungsmaschine fuer Uhren- und kleinmechanische Bestandteile . . .*».

Conclusion

Les problèmes de l'analyse automatique de textes sont donc très nombreux. Ils sont d'ordre linguistique, sémantique, grammatical, orthographique, etc. Si pour une documentation scientifique et technique la méthode des codes de classification et des mots-clés attribués manuellement par l'homme est suffisante, pour une documentation économique, juridique ou politique, cela ne suffit plus. Une analyse du texte est nécessaire, car ce type de document a été rédigé généralement avec beaucoup de soins, où chaque mot a été pesé et où la place d'une virgule possède une importance immense, comme dans le cas de traités économiques ou politiques, résultats d'un très long travail diplomatique. C'est pourquoi il est probable que l'analyse automatique de textes sera de plus en plus utilisée, malgré les problèmes (d'ailleurs en bonne partie résolus) soulevés par cette technique.

Le problème de(s)	est résolu par la technique des	utilisée dans les systèmes
Différentes formes grammaticales	Masques	*DPS, Text-Pac, Stairs*
Signes de ponctuation et signes diacritiques	Tests de caractères spéciaux	*Typesetting, Edit*
Notions ou mots composés	Positions relatives des mots dans une phrase	*DPS, Text-Pac, Stairs*
La reconnaissance de la langue	Prépositions	En développement
La synonymie	Chaînages	*IRMS, Stairs*
L'indexation à l'aide de mots-clés ne figurant pas dans le texte	Structures et proximités sémantiques	En développement

L'utilisation de l'ordinateur représente la seule possibilité actuelle de faire face au flot d'informations et de papier qui envahit chaque bibliothèque, chaque centre de documentation, chaque individu dans sa vie professionnelle et privée. Malgré les difficultés et les problèmes, l'analyse automatique de textes, technique jeune et nouvelle, représentera une aide importante et permettra aux économistes, aux juristes et aux politiciens d'accomplir leur tâche encore mieux et de manière encore plus efficace.

MITTEILUNGEN VSB – COMMUNICATIONS DE L'ABS

AUS DER TÄTIGKEIT DES VSB-VORSTANDES

In dieser Rubrik soll künftig von Zeit zu Zeit über Probleme berichtet werden, mit denen sich der Vorstand der Vereinigung schweizerischer Bibliothekare in seinen Sitzungen befaßt und die für einen weiteren Kreis unserer Mitglieder von Interesse sein können.

An der Vorstandssitzung vom 23. März 1972 in Bern kam u. a. die Frage der «*International Standard Book Number*» (ISBN) zur Sprache. Sie stellt vorwiegend ein Instrument des Buchhandels dar; er hat im Grunde genommen die Zuteilung der ISBN vorgenommen. Ein Instrument des Buchhandels bildet jedoch — wenigstens vorläufig noch — auch das «Schweizer Buch», selbst wenn es zunächst, wie