Organisation, information, environnement social et économique

COMMENT CARACTERISER UN DOCUMENT TECHNIQUE A L'AIDE DE L'ANALYSE DE DONNEES

TECHNICAL DOCUMENT CHARACTERIZATION BY DATA ANALYSIS

94NO00011

EDF Electricité de France

Direction des Etudes et Recherches

EDF
Electricité de France

Direction des Etudes et Recherches

SERVICE INFORMATIQUE ET MATHÉMATIQUES APPLIQUÉES
Département Traitement de l'Information et Etudes Mathématiques

Mai 1993

MAUGET A.

COMMENT CARACTERISER UN DOCUMENT TECHNIQUE A L'AIDE DE L'ANALYSE DE DONNEES

TECHNICAL DOCUMENT CHARACTERIZATION BY DATA ANALYSIS

Pages : 14

94NO00011

Diffusion : J.-M. Lecœuvre
EDF-DER
Service IPN. Département SID
1, avenue du Général-de-Gaulle
92141 Clamart Cedex

ISSN 1161-0603

EXECUTIVE SUMMARY :

Nuclear power plants possess documents analyzing all the plant systems, which represents a vast quantity of paper. Analysis of textual data can enable a document to be classified by grouping the texts containing the same words. These methods are used on system manuals for feasibility studies. The system manual is then analyzed by LEXTER and the terms it has selected are examined.

We first classify according to style (sentences containing general words, technical sentences, etc.), and then according to terms.

However, it will not be possible to continue in this fashion for the 100 system manuals existing, because of lack of sufficient storage capacity. Another solution is being developed.

XIEN 860

Author

EDF
Electricité de France

Direction des Etudes et Recherches

SERVICE INFORMATIQUE ET MATHÉMATIQUES APPLIQUÉES
Département Traitement de l'Information et Etudes Mathématiques

Mai 1993

MAUGET A.

COMMENT CARACTERISER UN DOCUMENT TECHNIQUE A L'AIDE DE L'ANALYSE DE DONNEES

TECHNICAL DOCUMENT CHARACTERIZATION BY DATA ANALYSIS

Pages : 14

94NO00011

Diffusion : J.-M. Lecœuvre
EDF-DER
Service IPN. Département SID
1, avenue du Général-de-Gaulle
92141 Clamart Cedex

ISSN 1161-0603

SYNTHÈSE :

Les centrales nucléaires disposent de documents analysant tous les systèmes élémentaires de la centrale, ce qui représente une masse énorme de papier.

L'analyse de données textuelles permet de classifier un document, en regroupant les textes contenant les mêmes mots. Nous utilisons ces méthodes sur un dossier de système élémentaire, pour étude de faisabilité. Puis ce dossier est analysé par LEXTER, et nous étudions les termes sélectionnés par le code.

Dans le premier cas, nous classifions sur le style (phrases contenant des mots généraux, phrases techniques...). Dans le second cas, nous classifions sur les termes.

Ce travail ne pourra être conduit sur les cent dossiers de système élémentaire, faute de place mémoire. Une autre approche est donc en cours de développement.

EXECUTIVE SUMMARY :

Nuclear power plants possess documents analyzing all the plant systems, which represents a vast quantity of paper. Analysis of textual data can enable a document to be classified by grouping the texts containing the same words. These methods are used on system manuals for feasibility studies. The system manual is then analyzed by LEXTER and the terms it has selected are examined.

We first classify according to style (sentences containing general words, technical sentences, etc.), and then according to terms.

However, it will not be possible to continue in this fashion for the 100 system manuals existing, because of lack of sufficient storage capacity. Another solution is being developed.

plan

I- Les données étudiées: le document 2 d'un Dossier de Système Élémentaire (DSE).

II- Pourquoi informatiser?

III- Apport de l'analyse de données dans l'étude du texte origine.

IV- L'extraction des termes significatifs permet-elle d'améliorer la classification?

V- Retrouve-t-on les grands paragraphes du DSE?

I- Les données étudiées: le document 2 d'un Dossier de Système Élémentaire (DSE).

Un DSE (Dossier de Système Élémentaire) décrit une partie d'une centrale nucléaire suivant une structure fixe pour tous les éléments de la centrale.
On trouve 8 grands chapitres:
1-rôle du système
2-base de conception
3-description et définition des matériels
4-phénomènes physiques déterminant les conditions de fonctionnement
5-controle commande
6-évaluation de la sûreté
7-principe d'entretien et essais périodiques
8-protection contre les grands froids

II-Pourquoi informatiser?

Les DSE doivent avoir une structure identique, quel que soit le sujet traité. Cette structure est-elle bien adaptée aux documents écrits? Peut-on retrouver facilement le contenu en fonction du titre du chapitre? Peut-on regrouper les textes en fonction du sujet traité, des mots employés?
Sur un seul document, on peut faire l'analyse manuelle; Mais il y aura plus de 100 DSE représentant des dizaines de millier de pages.
On a donc besoin d'une méthode informatique pour caractériser, et classer, tous ces documents: Pour cela, on utilise l'analyse de données textuelles.

III- Apport de l'analyse de données dans l'étude du texte origine.

III-1 Le document étudié

Nous allons étudier le document 2 du DSE "Conditionnement de la salle de commande et des locaux annexes" (les documents 2 étant ceux qui contiennent le plus de texte).
Le document est mis en forme. C'est à dire que chaque paragraphe et sous-paragraphe a été séparé et est devenu un texte que nous numérotons. Nous obtenons 172 textes.
Voici le début d'un texte:

****A1
LE SYSTEME DVC A POUR ROLE :

DE MAINTENIR DES CONDITIONS D'AMBIANCE (TEMPÉRATURE ET HUMIDITÉ) QUI PERMETTENT LE BON FONCTIONNEMENT DES MATÉRIELS ET LE SÉJOUR DU PERSONNEL D'EXPLOITATION DANS LES LOCAUX DE L'ÎLOT DE SURVIE DÉSIGNES CI-APRÈS ET SITUES AU NIVEAU 15,12 M DU BÂTIMENT ÉLECTRIQUE, ENTRE LES FILES F H ET 2-9 :
. LA SALLE DE COMMANDE LC0919,

III-2 Classification des mots

Nous utilisons le logiciel d'analyse de données textuelles SPADT pour étudier ce document.
Ce DSE compte 20097 mots avec 2154 mots distincts. En sélectionnant les mots figurant plus de 10 fois, on retient 286 mots distincts. Comme on a gardé le texte intégral les mots les plus fréquents sont des mots outils comme : de, le, a...
Dans un premier temps, on supprime les mots outils. Mais cela pose des problèmes par la suite; en effet pour rechercher les groupes de mots comme "conditionnement d'air", si on supprime "d'" le logiciel SPADT ne sait plus regrouper "conditionnement" et "air". J'ai donc gardé tous les mots, ce qui ne pose pas vraiment de problème, car des mots utilisés dans tous les textes n'influencent pas l'analyse statistique.
De même, je n'ai pas lemmatisé les mots (mise en équivalence des singulier-pluriel, des verbes conjugués).

En faisant abstraction des mots outils, les termes les plus employés sont : voie (239 fois), air (190), iode (158), locaux (107)...

Ensuite on fait une analyse factorielle des correspondances des textes sur les mots sélectionnés. Puis on effectue une classification ascendante hiérarchique sur les textes caractérisés par leurs premières coordonnées factorielles. On garde 6 classes; Celles-ci sont caractérisées par la liste des mots les plus utilisés dans une classe.

classe 1 (69 textes)	locaux
	apports
	dans
	§
classe 2 (12 textes)	document
	annexe
	reporter (se)
classe 3 (61 textes)	batterie
	vanne
	piège
	ST
classe 4 (9 textes)	secteur
	essai
	feu
	détection
	incendie
classe 5 (1 texte)	RS
	KW
classe 6 (20 textes)	resp
	RS
	KG
	ZV
	enclenchement

On peut également chercher les phrases caractéristiques. Il y a 2 façons de les rechercher. Soit par fréquence d'utilisation; cette méthode retient les phrases plutôt courtes. Soit en sélectionnant les phrases les plus proches des classes de réponses au sens de la distance du CHI2; cette méthode favorise les phrases longues.
En faisant la synthèse entre les 2 méthodes on obtient :

classe 1: les phrases commençant par § et contenant le mot "locaux" (ex: reprise d'air dans les locaux...). Ce sont des phrases de description générale du système.
classe 2: "se reporter à l'annexe .. jointe au sous-document ..
classe 3: textes décrivant les files de conditionnement d'air, les files de filtration...
classe 4: description des tests (ex: défauts liés à la mise en essai d'un secteur de feu...).
classe 5: texte technique sur les batteries chaudes électriques.
classe 6: textes très techniques avec des descriptions précises contenant des numéros et des codes de matériel.

III-3 Classification des suites de mots.

Le logiciel SPADT permet également de travailler sur des suites de mots, que l'on appelle des segments répétés.

Voici quelques exemples de segments répétés:

filtration iode (57 fois)
file iode (30 fois)
file de conditionnement (12 fois)
centrale de conditionnement (15 fois)
air neuf (31 fois)
air neuf extérieur (9 fois)
armoire de climatisation (22 fois)

On fait une analyse factorielle des correspondances sur les segments répétés, puis une classification sur les textes caractérisés par leurs premières coordonnées factorielles. On garde également 6 classes; mais il ne s'agit pas à priori des mêmes classes que dans l'analyse précédente.
Les classes sont caractérisées par la liste des segments répétés les plus utilisés dans une classe.

classe 1 (52 textes)	dans les
	l'installation
	la salle
	îlot de survie
	installation centralisée
classe 2 (57 textes)	la centrale
	de filtration iode
classe 3 (8 textes)	secteur de feu
	du feu
	l'opérateur
classe 4 (16 textes)	la batterie
	batterie chaude électrique
	la vanne
classe 5 (24 textes)	105 RS
	103 RS
	RS et
	la file
	404 RA
classe 6 (15 textes)	se reporter
	l'annexe
	125 V
	au sous document

On retrouve facilement les réponses caractéristiques au vu des segments répétés les plus usités.

III-4: Faut-il classer sur des mots ou des groupes de mots?

Les textes sont-ils regroupés dans les mêmes classes avec les 2 méthodes?

mots / segments	1	2	3	4	5	6	
1	46	0	6	0	0	0	52
2	18	0	31	2	0	6	57
3	0	0	0	8	0	0	8
4	0	0	16	0	0	0	16
5	1	0	8	0	1	14	24
6	3	12	0	0	0	0	15
	69	12	61	9	1	20	

La classe 3 de la classification sur les mots se retrouve éparpillée lors de la classification sur les segments répétés. Sinon les autres classes se retrouvent assez bien.
Je pense qu'il est préférable de faire une classification sur les mots, car on est sûr de garder toutes les informations. En effet les mots utilisés ensemble se retrouvent proches; ils participeront donc ensemble à la classification. De même l'éloignement de 2 termes que l'on pensait proches peut être instructif.

IV- L'extraction des termes significatifs permet-elle d'améliorer la classification?

IV-1 Le document étudié

Le même DSE a également été analysé par le logiciel LEXTER. Ce logiciel permet de sélectionner les termes significatifs d'un document.
Voici le début du même document après analyse par LEXTER:

****A1
ACCÈS,
AMBIANCE,
BÂTIMENT,
BATIMENT_ELECTRIQUE,
BL,
BON_FONCTIONNEMENT_DES_MATERIEL,
BUREAU,
CONDITION_D_AMBIANCE,
CONTAMINATION,
CONTAMINATION_RADIOACTIF,
CONTAMINATION_RADIOACTIF_DU_SITE,

IV-2 la classification

Le logiciel SPADT n'autorise pas les mots dépassant 20 lettres . Nous décidons que les mots dépassant 19 lettres seront tronqués.

Les 174 textes contiennent 5099 mots (1313 mots distincts).
Les textes sont beaucoup plus riches que précédemment. En sélectionnant les mots figurant plus de 5 fois on obtient ainsi 215 mots distincts. Les mots les plus fréquents sont :

- file_de_conditionne (86 fois)
- conditionnement (79 fois)
- conditionnement_d_a (76 fois)
- piège (54 fois)
- batterie_chaud (46 fois)

Nous avons un problème pour l'analyse statistique des textes "Lextérisés". En effet LEXTER décompose les expressions. Par exemple on trouve "air_neuf_extérieur" et "air_neuf". Les mêmes mots seront donc comptés plusieurs fois, puisque pour SPADT ce sont des mots différents.

On utilise la même méthode statistique que précédemment, c'est-à-dire une analyse factorielle des correspondances, puis une classification sur les textes croisés par les termes de LEXTER.
On fait également une coupure en 6 classes, puis on cherche les mots caractéristiques des classes

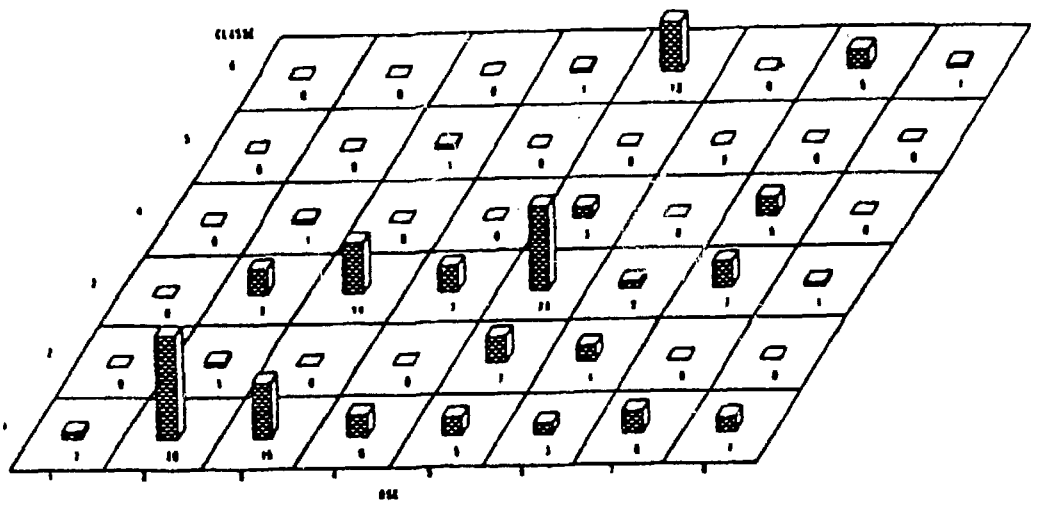
classe 1 (15 textes):	niveau_15
	couloir
	12_m
	bl
	débit_de_surpression
classe 2 (70 textes):	contamination_radio
	kg
	centrale_de_filtration
	filtration_iode
	registre_d_isolement
classe 3 (23 textes):	ttle
	essai_périodique
	détection_incendie
	apparition
	automatisme
classe 4 (56 textes):	climatiseur
	armoire
	ci
	climatisation
	batterie_chaud
classe 5 (4 textes):	KW
	puissance
	ventilateur_d_extraction
	sûreté
	RF
classe 6 (6 textes):	sous_document_2
	diagramme_fonctionnement

V- Retrouve-t-on les grands paragraphes du DSE?

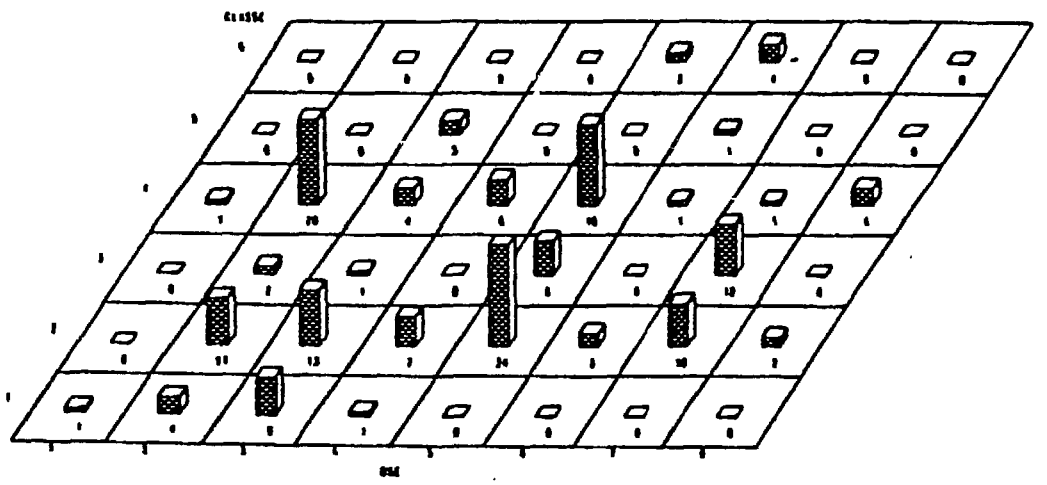
Le document 2 du DSE est structuré en 8 grands chapitres:
1-rôle du système (textes 1 et 2)
2-base de conception (textes 3 à 39)
3-description et définition des matériels (textes 40 à 69)
4-phénomènes physiques déterminant les conditions de fonctionnement (textes 70 à 83)
5-controle commande (textes 84 à 136)
6-évaluation de la sûreté (textes 137 à 145)
7-principe d'entretien et essais périodiques (textes 146 à 168)
8-protection contre les grands froids (textes 169 à 174)

Nous représentons la répartition des classes et des chapitres par des histogrammes.
Que ce soit au niveau des textes origines et des textes "lextérisés", on ne retrouve pas les chapitres du DSE.
Les classes issues de l'analyse sur le texte origine(sur les mots) caractérisent le style d'écriture (générale, technique...); Les classes du texte "lextérisé" caractérisent les termes techniques utilisés. Alors que dans les chapitres d'un DSE, on a aussi bien des termes généraux que des termes techniques, avec un style indépendant du chapitre.

repartition des textes par chapitres et par classes
textes originaux



repartition des textes par chapitres et par classes
textes issus de LEXTER



Conclusion

Il serait intéressant de refaire le même travail sur plusieurs DSE, pour chercher une classification entre les DSE. Mais nous serons vite limités par la taille mémoire nécessaire. Il semble préférable de s'orienter vers une autre méthode statistique. Nous allons prendre les termes sélectionnés par LEXTER, et rechercher des tests statistiques de proximités.

EDF
Electricité de France

Direction des Etudes et Recherches

Service Information Prospective et Normalisation

Département Systèmes d'information et de documentation

Groupe Exploitation de la Documentation Automatisée

1, avenue du Gal de Gaulle
92141 CLAMART Cedex
tel : 47 65 56 33

CLAMART Le 06/03/95

CEA
MIST/SBDS/SPRI
CENTRE DE SACLAY

91191 GIF SUR YVETTE CEDEX

à l'attention de :

MEMOIRE TECHNIQUE ELECTRONIQUE

****
Cette feuille est détachable grâce à la microperforation sur le coté droit.
****

Référence de la demande : **F514248**
Votre commande :
Origine : **CATALOGUE DES NOTES DER**

Numéro du document : **94NO00011**

Titre : **COMMENT CARACTERISER UN DOCUMENT TECHNIQUE A L'AIDE DE L'ANAL DE DONNEES**

Auteurs : **MAUGET A.**

Source : **COLL. NOTES INTERNES DER. ORGANISATION, INFORMATION, ENVIRONNEME**
Serial :

Référence du document : **SANS**

Nombre de pages: **0015** Nombre d'exemplaires : **001** Support : **P**