Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 70, No. 1 (2007) 3–26

Standardizing formats of corporate source data

CARMEN GALVEZ, FÉLIX MOYA-ANEGÓN

Scimago Research Group, Department of Information Science, University of Granada, Granada (Spain)

This paper describe an approach for improving the data quality of corporate sources when databases are used for bibliometric purposes. Research management relies on bibliographic databases and citation index systems as analytical tools, yet the raw resources for bibliometric studies are plagued by a lack of consistency in fied formatting for institution data. The present contribution puts forth a Natural Language Processing (NLP)-oriented method for the identification of the structures guiding corporate data and their mapping into a standardized format. The proposed unification process is based on the definition of address patterns and the ensuing application of Enhanced Finite-State Transducers (E-FST). Our procedure was tested on address formats downloaded from the INSPEC, MEDLINE and CAB Abstracts. The results demonstrate the helpfulness of the method as long as close control of errors is exercised as far as the formats to be unified. The computational efficacy of the model is noteworthy, due to the fact that it is firmly guided by the definition of data in the application domain.

Introduction

The general aim of this paper is to present a procedure for improving the data quality of the institutional address field when using databases. Data quality is a complex concept governed by multiple dimensions (completeness, correctness, currency, interpretability, and consistency) and may even depend on a number of rather subjective

Received February 3, 2006

Address for correspondence:
CARMEN GALVEZ
Scimago Research Group, Department of Information Science
University of Granada, 18071 Granada, Spain
E-mail: cgalvez@ugr.es

0138–9130/US $ 20.00

variables, often influenced by the context where data are used and also by specific users within a given context (CATARCI, 2004). In the case of research evaluation, databases provide information that is essential for bibliometric purposes. The results of quantitative studies are known to be determined by the quality of data, both within and across databases; yet quality control of data is still an issue (SHER et al., 1966; HAWKINS, 1977; 1980; GARFIELD, 1979; 1983a, b; WILLIAMS & LANNOM, 1981; PITERNICK, 1982; STEFANIAK, 1987; ANDERSON et al., 1988; LEYDESDORFF, 1988; MOED & VRIENS, 1989; DE BRUIN & MOED, 1990; BOURKE & BUTLER, 1996; INGWERSEN & CHRISTENSEN, 1997; HOOD & WILSON, 2003; VAN RAAN, 2005).

Many problems arise from the fact that most databases are primarily designed for the purpose of information retrieval, but not for secondary use, as in informetric research (HOOD & WILSON, 2003). Something similar can be seen in digital libraries, where these document repositories are used as a platform for bibliometric research (CUNNINGHAM, 1998). The main limitation commonly seen in conjunction with the use of databases is a lack of unification, with one same 'object' having different names (MOED, 1988). While this phenomenon can appear within a single database, it is even greater if a number of databases are merged (BRAUN et al., 1995; FRENCH et al., 2000). The principle shortcomings for bibliometric applications stem from:

1. Lack of consistency in author, journal, and institution names.
2. Lack of consistency in field formatting for author, journal, and institution data.

The present study is dedicated to the second problem,[*] and aims to develop a procedure that resolves the lack of uniformity in field formats, with an eye to establishing regularity in the way name data are structured in corporate sources. Such tools are indeed key for smoothing out the process of normalizing author affiliation for bibliometric analyses, because corporate source data of poor quality have enormous repercussions for collaboration indicators, the delimitation of scientific fields and evaluative scientometrics. In a continued effort to improve the bibliometic data-system, the Leiden Centre for Science and Technology Studies (CWTS) has developed a computerized and manual procedure for cleaning up bibliographic data and unifying hierarchical structures addresses by using the CWTS thesaurus of main organizations and their scientific addresses database. A detailed description of the CWTS data-system is given in MOED et al., (1995).

Many researchers depend on corporate addresses in view of the increasing impact of studies about research that are centred on institutional domains (CARPENTER et al., 1988; MOED & VAN RAAN, 1988; SHRUM & MULLINS, 1988; DE BRUIN & MOED,

[*] The first problem, relative to the lack of consistency in institutional names, has been dealt with in an earlier study submitted to *Scientometrics* and now under review by editor and referees (GALVEZ & MOYA-ANEGÓN, in revision).

1993; RINIA et al., 1993; HERBERTZ & MÜLLER-HILL, 1995; MELIN & PERSSON, 1996; BOURKE & BUTLER, 1998; VAN DEN BERGHE et al., 1998; NOYONS et al., 1999; MÄHLCK & PERSSON, 2000; MOED, 2000; MOYA-ANEGÓN et al., 2004). Although manual processing will, to some extent, be inevitable for unifying and reformatting the institutional affiliations of authors, it is hoped that the new approach described here provides a means of overcoming the scattering of organization data, thereby facilitating data isolation and unification for later bibliometric analyses.

The problem of data formatting in corporate sources

Scientific publications always enclose key data regarding author affiliation, information which is to be processed by bibliographic database and citation index systems. One of the pitfalls soon encountered along these unification processes lies in the arbitrary manner in which the name data are structured in the address field of the file. Authors do not use a standard code for affiliation data in scientific publications; though databases such as Science Citation Index Expanded (SCI-E), the Social Science Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI) of the Institute for Scientific Information (ISI–Thomson Scientific, Philadelphia, US) afford some logical or hierarchical order in addresses – for instance, names of universities are placed on the whole at the beginning of the address (DE BRUIN & MOED, 1990). The corporate source field in the ISI databases consists of several items separated by commas and semicolons: the first parts of the address refer to the organization and usually contain the names of overall organizations such as universities or hospitals, divisions such as schools or faculties, and subdivisions such as departments or sections (DE BRUIN & MOED, 1993).

Although it is widely accepted that the two final elements in corporate sources indicate the city and country where the organization of reference is located, the number of parts used to define this name can differ substantially. And so, raw publication data contain much needless variation in reporting the institutional name. This variety of formats in the address fields results in an eventual “scattering” of affiliations, interfering with the recognition and isolation of the data set regarding a particular organization or subdivision for subsequent bibliometric analyses. Not only does this problem arise within a given database; indeed, it is compounded if we try to gather up data from different databases. For an initial appraisal of this situation, we offer the reader some examples of structural differences in address formats from the MEDLINE, SCOPUS, CAB Abstracts and ISI Web of Science (ISI-WOS) databases, here in Table 1.

Table 1. Several formats of corporate affiliations

Databases	Data formatting of the address field
MEDLINE	DEPARTMENT OF MOLECULAR BIOTECHNOLOGY, GHENT UNIVERSITY, COUPURE LINKS 653, 9000 GENT, BELGIUM
	GHENT UNIVERSITY, DEPARTMENT OF MOLECULAR BIOTECHNOLOGY, COUPURE LINKS 653, 9000 GENT, BELGIUM
CAB Abstracts	DEPARTMENT OF MOLECULAR BIOTECHNOLOGY, FACULTY OF AGRICULTURAL AND APPLIED BIOLOGICAL SCIENCES, GHENT UNIVERSITY, GENT, BELGIUM
	DEPARTMENT OF MOLECULAR BIOTECHNOLOGY, GHENT UNIVERSITY, COUPURE 653, B-9000 GENT, BELGIUM
SCOPUS	DEPARTMENT OF MOLECULAR BIOTECHNOLOGY, GHENT UNIVERSITY, BELGIUM
	GHENT UNIVERSITY, DEPT. OF MOLECULAR BIOTECHNOLOGY, 9000 GENT, B
	MOLECULAR BIOTECHNOLOGY DEPARTMENT, FAC. OF AGRIC. AND APPL. BIOL. SCI., GHENT UNIVERSITY, GHENT, BELGIUM
ISI-WOS	STATE UNIV GHENT, DEPT MOL BIOTECHNOL, GHENT, B-9000 BELGIUM
	STATE UNIV GHENT, FAC AGR & APPL BIOL SCI, DEPT MOL BIOTECHNOL, GHENT, B-9000 BELGIUM

The inconsistencies in formatting make it very difficult to automatically parse this field into its constituent parts. Some handcrafted post-processing is needed to ensure order and consistency in maneuvering corporate source data. This situation led us to delve into two main objectives: 1) the development of a procedure that would allow for the tagging of this type of sequence; and 2) the application of some type of automatic process to help us to recognize equivalent structures and unify them in a fixed format.

It is important to point out that beyond the scope of the present work remain those problems originating in any errors or inconsistencies produced by abbreviations, transliteration differences, differences in spelling, or name changes. Nor do we tackle problems deriving from the absence in the address of the first institutional level, or difficulties in the assignment of each document to a center that may result from ambiguity or inconsistency in the use of different names to refer to a single institution, cases where a single same name may designate two or more separate institutions, or assigneeship reflecting different nationalities. The validation and correct institutional assignment of addresses is a task corresponding to experts.

Our proposal – bootstrapping and mapping structures with transducers

The *a priori* understanding guiding our initiative is that the affiliation addresses of scientific publications can be considered structured entities, even if their structure is not manifest. A Named Entity (NE) is a sequence of words that refers to an entity such as persons or organizations. The problem with NE recognition would be a task corresponding to Information Extraction (IE), with IE defined as the set of techniques and methods used to obtain structured data from natural language texts (HOBBS, 1993). In IE processes, texts are taken as input in order to produce fixed formats or unambiguos data as the output. This data may be used directly for display to users, may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in Information Retrieval (IR) applications (CUNNINGHAM, 2005).

Information Extraction technology arose in response to a need for efficient processing of texts in specialized domains. It focuses only on the relevant parts of the text, disregarding the rest (GRISHMAN, 1997). As in any other discipline pertaining to Natural Language Processing (NLP), one of two basic approaches may be adopted: linguistic (JACOBS & RAU, 1990; HOBBS et al., 1992; ABNEY, 1996) and statistical (NERI & SAITTA, 1997). Linguistic techniques are based on a specialized corpus, and lexical and knowledge resources (such as dictionaries, regular expressions, or patterns and grammars) developed to identify the information to be extracted in the target domain. Statistical techniques, on the other hand, are based on the use of a corpus of data pre-annotated according to the information to be extracted and automatic learning methods. The choice of orientation should be guided by the specific means at our disposal: a knowledge-based approach is chosen if we have lexical and knowledge resources and an un-annotaded corpus; whereas a statistical approach is preferred if we have an annotated corpus (WATRIN, 2003).

These are the two basic options; however, we face the major impediment of not having the lexical resources that would allow us to tag this type of entity, nor pre-annotated data that would allow for a more automatic learning process of the resources. To overcome these two shortcomings we propose a 'hybrid' approach: a knowledge based extraction procedure in which manually pre-annotated patterns are defined, and therefore likely to target all the information to be extracted. This general proposal will be developed in the following stages: (i) Definition of corporate address patterns in terms of constituent analysis; (ii) Address matching and bootstrapping structures, to recognize the address patterns, then classify and mark the parts of addresses associated with the corresponding structures; and (iii) Mapping structures, establishing the transformational operations that will allow us to map equivalent structures onto a standardized structure or common format.

Definition of address patterns

In order to identify the structured patterns of corporate source data, we shall first adopt what is known as Immediate Constituents (IC)* analysis, a well-known method in linguistics, based on the notion that between sentences and words there exist a series of intermediate degrees with a hierarchical order that divides sentences into successive layers, or constituents, until arriving at the final layer. The purpose of IC analysis would be to determine and show the interrelations between words in a given linguistic structure.

* The term Immediate Constituents analysis was introduced by American structuralists through the application of formal methods of linguistic analysis. CHOMSKY (1957) made the first significant technical contribution to linguistics by formalizing Immediate Constituent analysis by means of Context Free Grammars.

Under IC analysis, these patterns would be organized in sets, with consituent labels. We thus define address patterns as sequences of constituents separated by a delimiter character – most often a comma – along with some of the constituents containing triggering words, or 'core terms', that define the nuclei of the address pattern (such as 'Department', 'Faculty', 'University', 'College', 'Institute', 'School', 'Clinic', 'Centre', 'Hospital', 'Laboratory', 'Foundation' or 'Group'). These triggering words are activated within a specific context and serve as selectional restrictions* – which is a way of handling, in linguistics, the free order of the constituents, and is applied to resolve structural ambiguities. For instance, in an address downloaded from the SCOPUS database we can distinguish five immediate constituents:

[DEPT. OF MOLECULAR BIOTECHNOLOGY], [FAC. OF AGRIC. AND APPL. BIOL. SCI.], [GHENT UNIVERSITY], [B-9000 GHENT], [BELGIUM]

We might give these immediate constituents the labels *A* (**DEPT**. OF MOLECULAR BIOTECHNOLOGY), *B* (**FAC.** OF AGRIC. AND APPL. BIOL. SCI '), *C* (GHENT **UNIVERSITY**), *D* (B-9000 GHENT), *E* (BELGIUM). However, the description of the linear structure of this address would offer only the horizontal succession of the elements that make it up; and so only a hierarchical placement of the elements would reveal the relationships among the constituents. The tree-diagram (Figure 1) given below is to be read as follows: the ultimate constituents of the address pattern (such as the words 'Dept', 'Molecular', or 'Biotechnology') would, in turn, be the immediate constituents of a complex form indicated by node *A*.

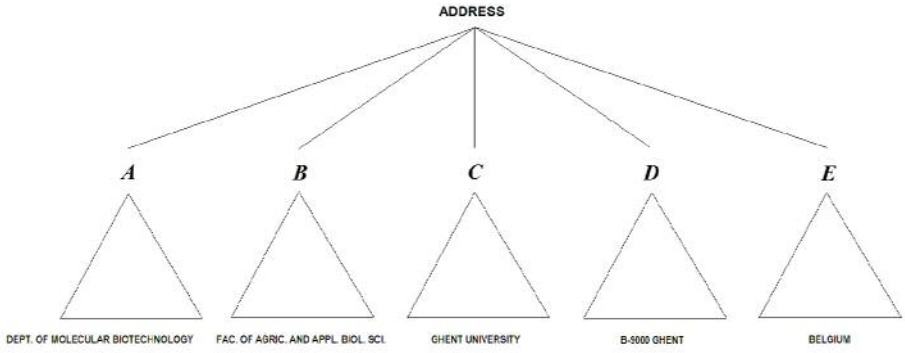


Figure 1. Tree diagram of the constituents of the address pattern

* Term coined by CHOMSKY (1965) to account for the variable order of syntactic structures. It is a formal device that limits the combinability of lexical units. The selectional restrictions imply a semantic selection to deal with the free order of constituents; they are usually an effective strategy in the case of very restricted domains, as is the case at hand.

Unfortunately, determining the boundaries of address components – a trivial assignment for humans – is difficult to model in automatic form. Like any other natural language processing task, it calls for a means of tagging this type of sequence. Therefore, the main problem is that the entities to be identified are unlabeled data, meaning that we must resort to some procedure for entity tagging in order to identify the structure and classify the component parts.

Address matching

After defining the address patterns in terms of IC and selectional restrictions based on triggering words that will help identify the relevant information, pattern-matching will be undertaken, using finite state techniques. We choose, from within this array of techniques, to apply finite automata and transducers. A finite automata accepts a string or a sentence if it can trace a path from the initial state to the final state by jumping along the stepping stones of labeled transitions. A finite automata is thus defined as a network of states and transitions, or edges, in which each transition has a label (ROCHE, 1996). Formally, a Finite-State Automata (FSA) is a tuple $\tau = <\Sigma, Q, q_0, F, \delta>$ where:

- Σ is the input alphabet
- Q is a finite set of states
- q_0 is the initial state, $q_0 \in Q$
- F is the final state, $F \subseteq Q$
- δ is a function of transition, $\delta: Q \times \Sigma \rightarrow Q$

To determine whether a string or sequence belongs to the regular language accepted by the FSA, the automata reads the string from left to right, comparing each one of the symbols of the sequence with the symbols tagging the transitions. If the transition is tagged with the same symbol as the input chain, the automata moves on to the following state, until the sequence is recognized in its entirety by reaching the final state. However, a finite automata is not capable of marking the parts of this sort of complex pattern.

One possible solution would be to develop a tagger for this type of sequence – though this would be very costly – then apply machine learning techniques,* in which learning algorithms take on a corpus of un-annotated sentences as input and return a corpus of bracketed sentences (VAN ZAANEN, 1999); this type of algorithm is used

* Language learning algorithms can be divided into two main groups, supervised and unsupervised ones, depending on the amount of information about language they use (VAN ZAANEN, 1999). The learning process of these algorithms consists of receiving, as input, several examples described by a set of attributes with its corresponding class label (these examples are the training set); then, the learner uses this training set to construct a hypothesis that will help it classify new instances.

more and more frequently for the boostrapping structure in natural language applications. The term 'bootstrapping refers to problem setting in which one is given a small set of labeled data and a large set of unlabeled data, and the task is to induce a classifier' (ABNEY, 2002, p. 360). The lack of resources and dictionaries for annotating the named entities, along with the need to recognize extraction patterns in specific domains from an untagged corpus, have led to the proliferation of bootstrapping methods.

In our proposal, a finite-state method is applied to the problem of bootstrapping structures. We adopt a simplified conception of bootstrapping methods in order to recognize and classify sequence chunks that represent corporate addresses. For this purpose we shall redefine, for the sake of convenience, the notion of bootstrapping as: a problem in which one is given a set of patterns with internal variables that mark the component parts (that is, its structure) and a large set of unannotated data, and the task will likewise be to induce a classifier. In this abridged approach, the proposal for bootstrapping structures will be based on the use of transducers. Here, the procedure will assign a structure to the corporate address that resembles the human-performance-type structure most appropriately given to these sequences.

A Finite-State Transducer (FST) is just like an FSA, except that the transitions have both an input label and an output label. An FST transforms one string into another string if there is a path through the FST that allows it to trace the first string using input labels and, simultaneously, the second string using output labels. One outstanding class of FSTs are the Enhanced Finite-State Transducers (E-FST), defined as transducers that use internal variables to identify and position parts of recognized sequences (SILBERZTEIN, 1999). These variables are set during parsing: they can store affixes of matching sequences, and the contents can then be copied to the output part of the transducers.

Using the graphic interface known as FSGraph (SILBERZTEIN, 2000), we drew E-FSTs that would represent the possible structures of addresses, accounting for triggering words and delimiters. Each constituent is tagged by parentheses (to enter parentheses around the parts, we use the tag $ to indicate the output of the transducer). This procedure will suffice to identify and classify the parts that constitute the linear structures of institutional address patterns (Figure 2).

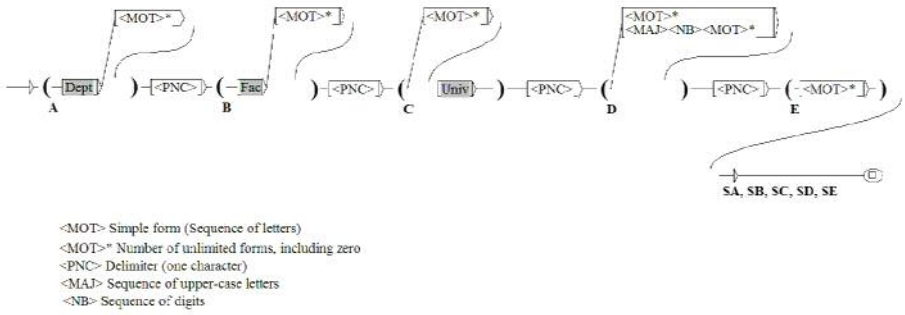


Figure 2. Linear structure of institutional address pattern

Mapping structures

Nonetheless, upon segmenting the address patterns we find that the IC order is not fixed – rather, it depends in many cases on database conventions, often determined by the variability of these data themselves within scientific publications. Thus, the relative positions of the constituents vary, giving rise to multiple alignment properties or structures in corporate data. Some possible combinations are:

B, A, C, D, E
C, B, A, D, E
A, C, D, E

In the face of this problem, we set forth: Let *A, B, C, D, E* be five constituents of some address pattern. *A* is said to be discontinuous, among other factors, if *A* is linearly ordered between *B* and *C*. Known as discontinuous are those constituents that are not found one beside the other, owing to different conventions. In the tree-diagram representation there will be intersections of the branches. Therefore, a syntagmatic and strictly superficial processing would be inadequate for dealing with the variety of possibilities in constituent order, which would notwithstanding give rise to equivalent structures.

In order to elaborate a standardized format with a fixed alignment we will need to develop a procedure in charge of mapping surface structures onto regularized structures. To this end we adopt a transformation method based on the interchangeability of constituents, in following the idea of American structuralist HARRIS (1951), according to whom constituents of the same type can be replaced by each other. Under the transformational theory of HARRIS (1951) we have as independent operations permutation, addition, substitution, adjunction, conjunction and suppression of constants. The value of the transformational method in light of our objectives resides in its capacity for detecting equivalent structures and producing uniform structures in institutional addresses.

If address structures can be subjected to the same transformational procedure, we infer they are identically structured; but if they cannot, their structure is different. From our viewpoint, some of the operations of transformation could then be modeled using E-FST.* The use of variables in transducers allows us to perform the relevant modifications in texts (SILBERZTEIN, 2000):

- *Erasure elements*: the replacement of *A B C* by *A C* allows us to erase the sequence stored in memory.
- *Insertions*: the replacement of *A B C D E* by *A B Univ C D E* allows us to insert the text '*Univ*' between sequences *B* and *C*.
- *Duplications*: the replacement of *A B C D* by *A A B C D* allows us to copy sequence *A* at two locations.
- *Permutations*: the replacement of *A B C D* by *C A B D* allows us to change the respective positions of *A* and *C*.

For the time being, we shall limit our focus to the interchangeability of constituents performed through permutation transformations. Using the same graphic interface, we drew enhanced transducers to represent the possible structures of addresses, able to produce as output a fixed-format preselected as the standardized form (Figure 3). In this case we decided to sort by constituent permutation, moving the main organization to the initial position.

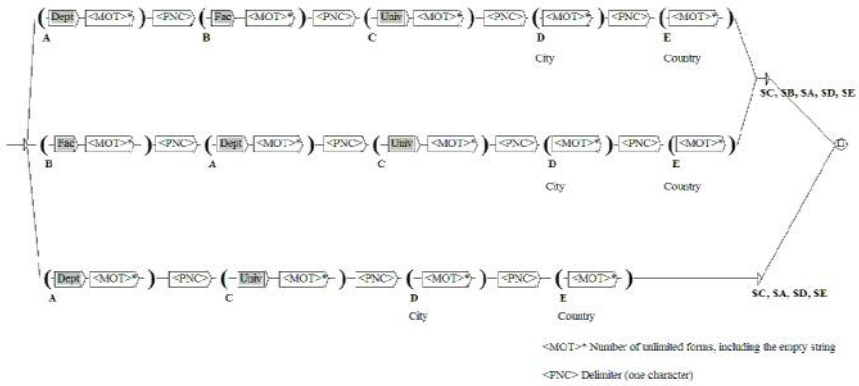


Figure 3. Graphical scheme of producing standardized form

* Enhanced transducers use internal variable to identify and place parts of recognized sequences. This function is similar to one carried out within programs like UNIX type SED (SILBERZTEIN, 2000). SED, AWK, and PERL are some of the UNIX utilities that implement *Regular Expressions*, mostly in tasks requiring pattern matching and substitution. SED is a **S**tream **Ed**itor, which follows commands just like an interactive editor to perform repetitive *search-and-replace* commands untouched by the human hand.

The aim of these transformational operations is to explain the equivalency relations between structures of corporate addresses that have the same set of constituents, and so the constructions are transformed one into the other if – and only if – there is coincidence of the constituent parts and conditions of occurrence. After a pre-processing stage, the application of this transducer will bring the main organization into the first position in the structured format we have selected to represent corporate addresses:

GHENT UNIVERSITY, FAC OF AGRIC AND APPL BIOL SCI, DEPT OF MOLECULAR BIOTECHNOLOGY, B-9000 GHENT, BELGIUM

By applying E-FST to the data structure of corporate addresses, we manage: (i) to reveal equivalences and differences in the structure of the units being examined; (ii) to expose the structural potential of the unit that will provide for sorting and classifying the parts of corporate sources; and, most importantly, (iii) to avoid any substantial modification of the corporate source data indicated by the author/s that might trigger greater problems in posterior quantitative analyses.

Methodology

We shall describe the components of corporate sources in tems of IC, which for possible bibliometric applications will be considered independent units of analysis (UA). Afterwards, we proceed to identify and structure these data using finite-state graphs compiled in transducers. The implementation of such a process is straighforward: first, the recognition process is activated because the sentences contain core terms; second, the address pattern is matched against the sentences and so the components of address patterns will be identified, labeled and permuted. This procedure will eventually allow us to establish an equivalence relation through permutation transformations, enabling us to produce organization names with a standard position for their components.

The application domain data

Because we use a corpus-based methodology, our system begins with a training phase, during which we learn the relevant address patterns from the application domain. To design the model, we took a sample of corporate names from INSPEC (produced by the Institution of Electrical Engineers), MEDLINE (U.S. National Library of Medicine), and CAB Abstracts (CAB International) databases, all in online version. The choice of these databases is justified by the fact that they no contain uniform format for corporate sources and require manual post-processing to clean-up, order and reformat these fields for automated bibliometric analysis. The dataset need not be very large, as there are

only so many legitimate structural forms of addresses, and so after a certain point, the larger the sample, the lesser the variations.

The training sample, downloaded from these databases, was gathered from organization names in higher education sectors with the terms 'Univ' and 'Dept' in the research address field. We limited the definition of address patterns to those containing these trigger words, because if no sort of restraint were set down, representation would be too extensive a task for testing the procedure. Moreover, this type of university address can nearly always be identified by certain well-known particles and abbreviations, whereas among data from other private sectors it is more difficult to find nucleus terms with clear borderlines. Below, Table 2 offers some examples of these corporate patterns.

Table 2. Excerpts of definitions of address patterns

Type	Address-Pattern	Instance
T_1	DEPT UNIV CITY CNTY	CISE Dept, Florida Univ, Gainesville, FL, USA
T_2	DEPT FAC UNIV CITY CNTY	Department of Applied Chemistry, Faculty of Engineering, Osaka University, Suita, Osaka 565-0871, Japan
T_3	DEPT DIV UNIV CITY CNTY	Department of Medicine, Division of Rheumatology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
T_4	DEPT LAB UNIV CITY CNTY	Department of Chemistry, Inorganic Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QR, UK
T_5	DEPT INST UNIV CITY CNTY	Department of Chemistry, Institute of Fundamental Sciences, Massey University at Auckland, New Zealand
T_6	DEPT GRP UNIV CITY CNTY	Department of Science & Technology, Applied Optics Group, University of Twente, Enschede, The Netherlands
T_7	DEPT UNIV CNTY	Department of Psychology, Bar-Ilan University, Israel
T_8	DEPT FAC UNIV	The Department of Pathology, Faculty of Medicine, Mansoura University
T_9	DEPT INST UNIV	The Department of Radiation Oncology, National Cancer Institute, Cairo University
T_{10}	UNIV DEPT CITY CNTY	University of Helsinki, Department of Public Health, Helsinki, Finland
T_{11}	FAC DEPT UNIV CITY CNTY	Veterinary Faculty, Department of Pathology, Akdeniz University, Eurdur, Turkey
T_{12}	CR DEPT UNIV CITY CNTY	Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA
T_{13}	SECT DEPT UNIV CITY CNTY	Section of Ecology, Department of Biology, University of Turku, 20014, Turku, Finland
T_{14}	SECT DEPT FAC UNIV CITY CNTY	Section of Parasitology, Department of Genome Sciences, Faculty of Medicine, Kobe University, Kobe 650-0017, Japan
T_{15}	GRP DEPT UNIV CITY CNTY	Group of Complex Fluids Physics, Department of Applied Physics, University of Almeria, 04120 Almeria, Spain
T_{16}	PROG DEPT UNIV CITY CNTY	Water and Watershed Research Program, Department of Biology, University of Victoria, Victoria, Canada
T_{17}	LAB DEPT UNIV CITY CNTY	Wave Phenomena Laboratory, Department of Physics, National Central University, Chungli, Taiwan 32054, Republic of China
T_{18}	LAB DEPT FAC UNIV CITY CNTY	Laboratory of Reaction Engineering, Dept of Chemical Engineering, Faculty of Engineering, University of Porto, Porto, Portugal
T_{19}	INST DEPT UNIV CITY CNTY	The Robert Hill Institute, Department of Molecular Biology and Biotechnology, University of Sheffield, UK
T_{20}	UNT DEPT UNIV CITY CNTY	Bird Ecology Unit, Department of Biological and Environmental Sciences, University of Helsinki, PO Box 65, Finland
T_{21}	DIV DEPT SCH UNIV CITY CNTY	Division of Bacterial Pathogenesis, Department of Microbiology, Graduate School of Medicine, University of the Ryukyus, Nishihara, Japan
T_{22}	DIV DEPT UNIV CITY CNTY	Division of Neonatology, Department of Pediatrics, Johns Hopkins University, Baltimore, USA

Units of analysis against constituent analysis

Given that the only standard convention of the address patterns is the separation of parts by means of delimiters (commonly commas), our initiative is to develop a model for the sorting of corporate source data with possible later bibliometric applications. We need to previously outline the address patterns in terms of independent units of analysis (UA) – defined as the objects of study (MCGRATH, 1996) in bibliometric research. The units of analysis will correspond to the Immediate Constituents (IC) in linguistic structures. We classify the components of address pattern structures as: UA1 (University), UA2 (Faculty/Hospital/Institute), UA3 (Department), UA4 (Centre/Unit/Section/Laboratory/Division/Research Group), UA5 (City); and UA6

(Country). In this way, address patterns such as 'DEPT UNIV CITY CNTY' will be redefined as 'UA3 UA1 UA5 UA6'.

Standardizing address formats via E-FST

E-FST, as we mentioned earlier, feature an input part, an output part and internal variables, the latter used during parsing to classify the parts of the sequences recognized. The use of internal variables lets us establish the order of the recognized sequences, and make any necessary permutations or insertions. It also allows us to intentionally modify the conditions of this synchronization to obtain the correct matching forms. Inputs and outputs are synchronized by means of the internal variables to store parts of the matching input sequence (tagged with the symbol $).

In order to determine a sorting of components of address patterns, we drew handcrafted finite-state graphs by means of an interface (SILBERZTEIN, 2000). The variables that we proposed for the arrangement of corporate names would correspond, for the purposes at hand, to the following units of analysis (UA):

Variable $UA1 (Main organization)
Variable $UA2 (Division)
Variable $UA3 (Subdivision-1)
Variable $UA4 (Subdivision-2)
Variable $UA5 (City)
Variable $UA6 (Country)

We then created three finite graphs that would represent the possible structures of corporate names for each database: INSPEC, MEDLINE and CAB Abstracts. The finite graphs were compiled in a transducer in charge of determining whether two expressions could be made equivalent via substitutions for variables. The application of transducers allows us to parse address pattern variants and realize permutations in order to obtain canonical sequences. Appendix 2 shows a simplified version of the graphic transducer 'INSPEC-Graph' in charge of identifying and standardizing the corporate address found in the INSPEC database. The input part of the graph contains:

- Internal variables whose function is to bracket the various UAs.
- Grey nodes that contain references to other graphs. For instance,the grey node labeled **Univ** in the graph 'INSPEC-Graph' encloses references to another imbedded graph of the same name in charge of representing and identifying all the possible variants of the core term 'University' (such as 'Univ', 'univ', 'University', 'Universiteit', 'Université', 'Universität', 'Università', 'Universitet', 'Universidad', or 'Universidade'). The same is true of the node **Fac** or **Hosp**. Meanwhile, the grey node labeled **Const1**, also in the graph 'INSPEC-Graph', encloses references to other imbedded

graphs of the same name that include special symbols written inside angles (such as <MOT> to represent and identify any sequence of simple forms separated by a space; or <NB> to represent and identify any sequence of digits).

- Delimiters (such as <PNC> in charge of identifying the delimiter characters that separate the address pattern constituents).

The output part of the graph contains the variables that represent the structures selected as the standardized formats, introducing generalized organization in the beginning of address patterns, such as:

$UA1, $UA2, $UA6
$UA1, $UA2, $UA5, $UA6
$UA1, $UA3, $UA6
$UA1, $UA3, $UA5, $UA6
$UA1, $UA2, $UA3, $UA5, $UA6

Performance evaluation

This approach was tested on three samples of institutional data from bibliographic records, downloaded from the INSPEC, MEDLINE, and CAB Abstracts databases. The dataset covers the period 2004, and a total of 4500 records randomly selected (1500 from each database) with the terms 'Univ or University' and 'Dept or Department' in the address fields (AD). The data of study were collected together in relation to an overall organization and a suborganization, because our intention was to have examples containing at least of two units of analysis for evaluation, one being the 'main organization'.

The set of references obtained was imported to a bibliographic management system, ProCite database (version 5.0), in order to automatically generate a list of variants that could be quantified in the evaluation and to eliminate the duplicate addresses, leading to a reduction to 3916 different addresses (1192 from INSPEC, 1416 from MEDLINE, and 1307 from CAB Abstracts databases). Before attempting analysis, the lists of institutional names were put through a series of transformations to allow them to be processed in the text-file format: the lists were segmented into sentences, and punctuation signs for abbreviations were eliminated because they would cause confusion with the delimiter character for the different units of analysis. The next step was to apply the finite graphs to the occurrences of the selected addresses. The process of matching can be carried out in one of three ways: 'shortest matches', 'longest matches', or 'all matches'. We opted for only 'longest matches'. In Appendix 1 an extract of the data obtained after application of 'INSPEC-Graph' is shown.

Our framework for assessing E-FST output revolves around two criteria: completeness and correctness. Recall (completeness of the model) would indicate the proportion of address patterns that are standardized with respect to a set of lists of evaluation. We shall define it *ad hoc* as the percentage of correct address patterns standardized over total possible address names susceptible of normalization. The measure of precision (correctness of the model) assesses the accuracy of transducers, and could be redefined as the ratio of correct address patterns standardized from among the total address patterns identified by the finite graphs. Thus, completeness and correctness in our context are similar to the concepts of recall and precision in information retrieval. The two measures were determined through the following equations:

$$\text{Recall } (R) = \frac{\text{Number of Correct Address Patterns Standardized}}{\text{Total Number of Possible Address Patterns}}$$

$$\text{Precision } (P) = \frac{\text{Number of Correct Address Patterns Standardized}}{\text{Total Number of Address Patterns Standardized}}$$

Likewise, we redefined the *F-measure* (VAN RIJSBERGEN, 1979), which stands for the harmonic mean of recall and precision (as compared to the arithmetic mean) and exhibits the desirable property of being highest when both recall and precision are high. Its calculation entails the following equations:

$$F_\beta = \frac{\left(\beta^2 + 1\right) RP}{\beta^2 R + P}$$

where the value of β controls trade-off:
β = 1: equal weight of recall and precision ($R = P$)
β < 1: weight of recall is higher
β > 1: weight of precision is higher

For the assessment of recall, precision and *F*-measure, we need the following frequency data:

- *Total number of possible address patterns*. To arrive at this figure, we identify the total number of address patterns that could be permuted to a standardized format. These data were obtained manually.
- *Total number of address patterns standardized*. To obtain this number, we took the total number of possible address patterns and subtracted the number of address patterns not standardized.

- *Number of correct address patterns standardized*. For these occurrences, we compared the transducer's output to its input and identified the address patterns that had been successfully standardized, removing under-standardization and over-standardization errors.

The under-standardization errors occur when address names are not reduced to an unified format, a type of error affecting recall. Over-standardization errors occur when address names are standardized incorrectly because they are not actual address patterns, or are non-valid patterns, an error affecting precision. The percentage of under-standardization and over-standardization errors could therefore be calculated as follows:

$$\text{Under-standardization Errors} = \frac{\text{Number of Address Patterns not Standardized}}{\text{Total Number of Possible Address Patterns}}$$

$$\text{Over-standardization Errors} = \frac{\text{Number of Non-valid Address Patterns Standardized}}{\text{Total Number of Address Patterns Standardized}}$$

Results and discussion

We shall now expound and analyze the results of applying the E-FSTs to the lists of corporate address, as shown in Tables 3 and 4. An analysis of the results in INSPEC shows the address patterns to be standardized with a high recall of R=0.99; and similarly solid results are seen in CAB Abstracts, with R=0.98. With respect to the value β=1, in INSPEC and in CAB Abstracts, we obtain the baselines F_1=0.99 and F_1=0.98, respectively. This is because the under-standardization rates for one and the other are 0.08 and 0.6, a relatively low proportion. In contrast, a fairly poor result of R=0.94 was obtained for MEDLINE, the failures in the coverage rate being 4.2, below the baseline F_1=0.96. In general, this deficiency of recall unchained by errors occurs because the E-FST cannot unify address patterns that are either not valid or else were not specified in the previous stage of definition of structures.

Table 3. Recall, precision and F-measure

	INSPEC	MEDLINE	CAB Abstracts
Possible address patterns	1192	1416	1307
Standardized address patterns	1191	1357	1299
Non-valid address patterns	1	83	14
Correct standardized address patterns	1191	1333	1293
Recall	0.99	0.94	0.98
Precision	1.00	0.98	0.99
F_1	0.99	0.96	0.98

Table 4. Error percentages

	INSPEC	MEDLINE	CAB Abstracts
Under-standardization errors	0.08	4.2	0.6
Over-standardization errors	0.00	1.8	0.5

The percentage of unmatched data derives from cases of non-valid addresses, some of which we show below (all taken from the MEDLINE database as it presented the greatest lack of recall):

- Errors in format arising from the lack of delimitation of the different parts of the institutional data (e.g., 'Department of Internal Medicine II Hokkaido University Graduate School of Medicine Sapporo Japan').
- The overlapping of constituents, causing a misrepresentation of formats, which usually stems from different components appearing joined, usually by a conjunction, as if they pertained to a single UA (e.g., 'Department of Chemistry **and** the Center for Nanofabrication and Molecular Self-Assembly, Northwestern University, Evanston, IL 60208-3113, USA').
- The misunderstanding of formats, produced by exceptions in components integrating the corporate names and which were not considered previously, when defining the address patterns to be extracted (e.g., '**Discipline of Microbiology and Immunology**, Department of Biosciences and Oral Diagnosis, School of Dentistry of Sao Jose dos Campos, Sao Paulo State University').
- Non-legitimate address patterns, caused by sequences that do not actually pertain to the institutional address (e.g., '**Professor and Associate Chair for Research**, Division of Cardiology, Department of Internal Medicine, Wayne State University, USA').
- The incapacity of the E-FST in mapping the institutional name to a uniform format when there are several candidates for 'core terms' (e.g., 'Department of Psychology, 1 **University** Station A8000, **University** of Texas, Austin, TX 78712, USA').

Analysis of results in terms of recall reveals, moreover, a problem of unmatching, in that it is not possible to achieve a complete formulation of all the structural phenomena in the addresses. As a practical application, our study never lost sight of the domain application data, which were used to define the address patterns, construct the model, and also to evaluate results; yet even so we were not able to know *a priori* precisely which structures we would encounter. The methodology would therefore require one to always account for a certain margin of structural uncertainty. When more format errors occur and the units of analysis are not separated by delimiters, results are poorer, as seen in the case of the MEDLINE database. This risk of non-standardization could be reduced through a handcrafted pre-processing stage: once data are downloaded, offline correction could modify the sequences that result in non-valid addresses (adding delimitation characters, separating the overlapping organization names, or eliminating strings that give rise to confusion).

In the precision phase of the experiment, the results in INSPEC give a particularly high precision of P=1, well above the baseline of F_1=0.99. The assessment in MEDLINE, with P=0.98, and CAB Abstracts, with P=0.99, were also very good, with F-measure scores respectively exceeding F_1=0.96 and F_1=0.98. Over-standardization is seen in a comparatively low proportion: 1.8 in MEDLINE, and 0.5 in CAB Abstracts. Most of these over-standardization errors occur because the transducer identifies and intends to unify some non-valid address patterns as if they were correct sequences (e.g., the non-legitimate address 'Professor of Neuroradiology, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands' is transformed to the equally non-legitimate 'Professor of Neuroradiology, Leiden University Medical Center, Department of Radiology, Leiden, The Netherlands'). This type of failure might be avoided altogether through the pre-processing and offline correction of downloaded data to amend the non-valid names, though corrections would more logically be made in a post-processing stage: having identified any erroneous data, a handcrafted unification process is undertaken.

The very high precision index we obtained can be explained by the fact that the application was guided by detailed information. That is, the corporate structures to be identified by the E-FST were clearly predefined in address patterns. Similarly, an explanation of computational efficacy could be that the model was directed and determined by the data that we intended to encounter in the domain of application, rather than depending on random identification. Besides, applying E-FSTs in the 'longest match' mode established priority of the longer sequences (as opposed to 'all matches'); and the fact that the sample consisted of real examples of institutional names containing trigger words made results more predictable.

Conclusions

In the realms of collaboration indicators, delimitation of scientific fields, and evaluative scientometrics, any lack of consistency in the transcription of institutional names becomes a critical issue for performance measurement. One major problem is rooted in the need for uniformity in the field formats of corporate source data. In order to isolate, analyze, and determine the arrangement of data while avoiding the scattering of organizations in bibliometric methods oriented at the level of institutions, it is essential to create resources that resolve this issue. This paper presents one novel means of improving data quality by eliminating inconsistencies in address field formatting, transforming address patterns under a uniform structure, and permuting the main organization into the primary position. We considered corporate data as entities and used an NLP-oriented method to capture structures as input, then produce the fixed-format as output.

On the basis of the experiments performed, two significant conclusions can be drawn. First, in the assessment of recall, we found under-standardization caused by non-valid corporate structures, because E-FST cannot unify corporate structures that are not specified in the stage of definition of address patterns, resulting in unmatched data. Second, the precision of the E-FST in mapping structures to common formats was very high, giving very few over-standardization errors; those that did occur can be traced to the device's processing of some non-valid addresses as if they were valid ones. Therefore, the greatest weakness of E-FST in mapping formats stems from non-valid structures, a situation calling for manual intervention through offline corrections. This could be done either in a handcrafted pre-processing stage to modify strings that would give non-valid address, or else in a post-processing correction stage.

To come to a close, we may suggest the usefulness of this descriptive method that provides a theoretical platform for the systematic unification of formats, and affirm its efficacy as long as: a) there are no errors in the formats (that is, non-valid formats) to be unified; b) the addresses contain 'core terms' that aid identification of the different units of analysis; and c) the different units of analysis are properly separated by delimiters. Notwithstanding, a major part of the computational efficacy could be justified by a feature inherent to the proposed formalism: it is guided by data foreseen to appear in the application domain.

References

ABNEY, S. (2002), Bootstrapping. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia.

ABNEY, S. (1996), Partial parsing via finite-state cascades. In: *Proceedings of the ESSLLI'96 Robust Parsing Workshop*. Prague, pp. 8–15.

ANDERSON, J., COLLINS, P. M. D., IRVINE, J., ISARD, P. A., MARTIN, B. R., NARIN, F., STEVENS, K. (1988), On-line approaches to measuring national scientific output: A cautionary tale, *Science and Public Policy*, 15 : 153–161.

BOURKE, P., BUTLER, L. (1996), Standards issues in a national bibliometric database: The Australian case, *Scientometrics*, 35 : 199–207.

BOURKE, P., BUTLER, L. (1998), Institutions and the map of science: Matching university departments and fields of research, *Research Policy*, 26 : 711–718.

BRAUN, T., BROCKEN, M., GLÄNZEL, W., RINIA, E., SCHUBERT, A. (1995), "Hyphenation" of databases in building scientometric indicators: Physics briefs, SCI based indicators of 13 European countries, 1980–1989, *Scientometrics*, 33 : 131–148.

CARPENTER, M. P., GIBB, F., HARRIS, J., IRVINE, J., NARIN, F. (1988), Bibliometric profiles for British academic institutions: An experiment to develop research output indicators, *Scientometrics*, 14 : 213–234.

CATARCI, T. (2004), Special issue on data quality in cooperative information systems (Editorial), *Information Systems*, 29 : 529–530.

CHOMSKY, N. (1965), *Aspects of the Theory of Syntax, Massachusetts Institute of Technology*, Cambridge, Massachusetts.

CHOMSKY, N. (1957), *Syntactic Structures*, Mouton, The Hague.

CUNNINGHAM, H. (2005), Information extraction, Automatic, *Enclyclopedia of Language and Linguistics*, 2nd ed. Elsevier, Oxford.

CUNNINGHAM, S. J. (1998), Applications for bibliometric research in the emerging digital libraries, *Scientometrics*, 43 : 161–175.

DE BRUIN, R. E., MOED, H. F. (1990), The unification of addresses in scientific publications. In: L. EGGHE, R. ROUSSEAU (Eds), *Informetrics 1989/90*. Elsevier Science Publishers, Amsterdam, pp. 65–78.

DE BRUIN, R. E., MOED, H. F. (1993), Delimitation of scientific subfields using cognitive words from corporate addresses in scientific publications, *Scientometrics*, 26 : 65–80.

FRENCH, J. C., POWELL, A. L., SCHULMAN, E. (2000), Using clustering strategies for creating authority files, *Journal of the American Society for Information Science and Technology*, 51 : 774–786.

GALVEZ, C., MOYA-ANEGON, F. (under revision), The unification of institutional addresses applying parametrized finite-state graphs (P-FSG), *Scientometrics*.

GARFIELD, E. (1979), *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*, John Wiley, New York.

GARFIELD, E. (1983a), Idiosyncrasies and errors, or the terrible things journals do to us, *Current Contents*, 2 : 5–11.

GARFIELD, E. (1983b), Quality control at ISI, *Current Contents*, 19 : 5–12.

GRISHMAN, R. (1997), Information extraction: Techniques and challenges. In: M. T. PAZIENZA (Ed.), *Information Extraction*. Springer-Verlag, Rome, pp. 10–27.

HARRIS, Z. S. (1951), *Methods in Structural Linguistics*. University of Chicago Press, Chicago.

HAWKINS, D. T. (1977), Unconventional uses of on-line information retrieval systems: On-line bibliometric studies, *Journal of the American Society for Information Science*, 28 : 13–18.

HAWKINS, D. T. (1981), Machine-readable output from online searches, *Journal of the American Society for Information Science*, 32 : 253–256.

HERBERTZ, H., MÜLLER-HILL, B. (1995), Quality and efficiency of basic research in molecular biology: A bibliometric analysis of thirteen excellent research institutes, *Research Policy*, 24 : 959–979.

HOBBS, J. R. (1993), The generic information extraction system. *In: Proceedings of the Fifth Message Understanding Conference (MUC-5)*. Morgan Kaufman, San Mateo, CA, pp. 87–91.

HOBBS, J. R., APPELT, D. E., TYSON, M., MABRY, B., ISRAEL, D. (1992), SRI international: Description of the FASTUS system used for MUC-4. In: *Proceedings of the Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann, pp. 268–275.

HOOD, W. W., WILSON, C. S. (2003), Informetric studies using databases: Opportunities and challenges. *Scientometrics*, 58 : 587–608.

INGWERSEN, P., CHRISTENSEN, F. H. (1997), Data set isolation for bibliometric online analyses of research publications: Fundamental methodological issues, *Journal of the American Society for Information Science*, 48 : 205–217.

JACOBS, P. S., RAU, L. F. (1990), SCISOR: Extracting information from on-line news, *Communications of the ACM*, 33 : 88–97.

LEYDESDORFF, L. (1988), Problems with the 'measurement' of national scientific performance, *Science and Public Policy*, 15 : 149–152.

MÄHLCK, P., PERSSON, O. (2000), Socio-bibliometric mapping of intra-departmental networks, *Scientometrics*, 49 : 81–91.

MCGRATH, W. (1996), The unit of analysis (object of study) in biblometrics and scientometrics, *Scientometrics*, 32 : 257–264.

MELIN, G., PERSSON, O. (1996), Studying research collaboration using co-authorships, *Scientometrics*, 36 : 363–377.

MOED, H. F. (1988), The Use of on-line databases for bibliometric analysis. In: L. EGGHE, R. ROUSSEAU (Eds), *Informetrics 87/88*. Elsevier Science Publishers, Amsterdam, pp. 133–146.

MOED, H. F. (2000), Bibliometric indicators reflect publication and management strategies, *Scientometrics*, 47 : 323–346.

MOED, H. F., DE BRUIN, R. E., VAN LEEUWEN, TH. N. (1995), New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications, *Scientometrics*, 33: 381–422.

MOED, H. F., VAN RAAN, A. F. J. (1988), Indicators of research performance: Applications in university research policy. In: A. F. J. VAN RAAN (Ed.), *Handbook of Quantitative Studies of Science and Technology*. Elsevier Science Publishers, Amsterdam, pp. 177–192.

MOED, H. F., VRIENSV, M. (1989), Possible inaccuracies occurring in citation analysis, *Journal of Information Science*, 15 : 95–117.

MOYA-ANEGÓN, F., VARGAS-QUESADA, B., HERRERO-SOLANA, V., CHINCHILLA-RODRÍGUEZ, Z., CORERA-ÁLVAREZ, E., MUNOZ-FERNANDEZ, F. J. (2004), A new technique for building maps of large scientific domains based on the cocitation of classes and categories, *Scientometrics*, 61 : 129–145.

NERI, F., SAITTA, L. (1997), Machine learning for information extraction. In: M. L. PAZIENZA (Ed.), *Information Extraction*. Springer-Verlag, Rome, pp. 10–27.

NOYONS, E. C. M., MOED, H. F., LUWEL, M. (1999), Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study, *Journal of the American Society for Information Science*, 50 : 115–131.

PITERNICK, A. B. (1982), Standardization of journal titles in databases (letter to the editor), *Journal of the American Society for Information Science*, 33 : 105.

RINIA, E. J., DE LANGE, C., MOED, H. F. (1993), Measuring national output in physics: Delimitation problems, *Scientometrics*, 28 : 89–110.

ROCHE, E. (1996), Finite-state transducers: Parsing free and frozen sentences. In: A. KORNAI (Ed.), *Proceedings of the ECAI 96 Workshop extended finite state models of language*. ECAI, pp. 52–57.

SHER, I. H., GARFIELD, E., ELIAS, A. W. (1966), Control and elimination of errors in ISI services, *Journal of Chemical Documentation*, 6 : 132–135.

SHRUM, W., MULLINS, N. (1988), Network analysis in the study of science and technology. In: A. F. J. VAN RAAN (Ed.), *Handbook of Quantitative Studies of Science and Technology*. Elsevier Science Publishers, Amsterdam, pp. 107–133.

SILBERZTEIN, M. (1999), Text indexation with INTEX, *Computers and the Humanities*, 33 : 265–280.

SILBERZTEIN, M. (2000), INTEX: An FST toolbox, *Theoretical Computer Science*, 231 : 33–46.

STEFANIAK, B. (1987), Use of bibliographic data bases for scientometric studies, *Scientometrics*, 12 : 149–161.

VAN DEN BERGHE, H., DE BRUIN, R. E., HOUBEN, J. A., KINT, A., LUWEL, M., SPRUYT, E., MOED, H. F. (1998), Bibliometric indicators of university research performance in Flanders, *Journal of the American Society for Information Science*, 49 : 59–67.

VAN RAAN, A. F. J. (2005), Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods, *Scientometrics*, 62 : 133–143.

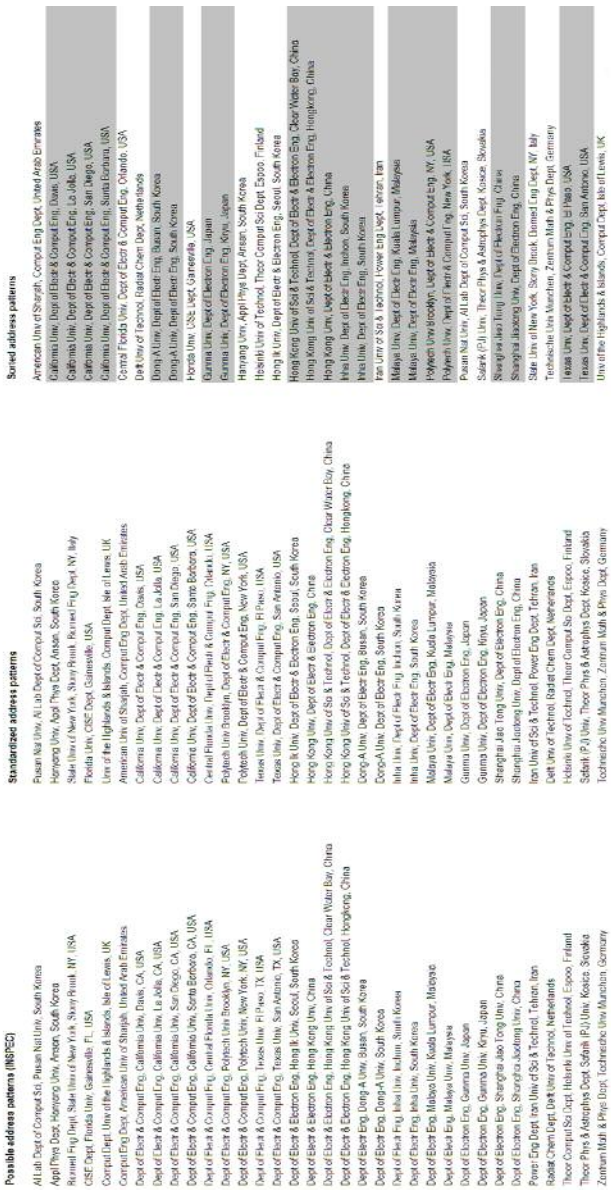
VAN ZAANEN, M. (1999). Bootstrapping structure using similarity. In: P. MONACHESI (Ed.), *Computational Linguistics in the Netherlands 1999-Selected Papers From the Tenth CLIN Meeting*. Universteit Utrecht, Utrecht, The Netherlands, pp. 235–245.

WATRIN, P. (2003), Information extraction and lexicon-grammar. In: *Proceedings of the Fourth Dutch-Belgian Information Retrieval Workshop*, DIR, Amsterdam, pp. 16–21.

WILLIAMS, M. E., LANNOM, L. (1981), Lack of standardization of the journal title data element in databases, *Journal of the American Society for Information Science*, 32 : 229–233.

Appendix 1

Excerpt of data obtained from the application of the INSPEC-Graph in a selection of address patterns found in the INSPEC database



Possible address patterns (INSPEC)	Standardized address patterns	Sorted address patterns
Ali Lab Dept of Comput Sci, Pusan Nat Univ, South Korea	Pusan Nat Univ, Ali Lab Dept of Comput Sci, South Korea	American Univ of Sharjah, Comput Eng Dept, United Arab Emirates
Appl Phys Dept, Hanyang Univ, Ansan, South Korea	Hanyang Univ, Appl Phys Dept, Ansan, South Korea	California Univ, Dept of Elect & Comput Eng, Davis, USA
Rommel Eng Dept, State Univ of New York, Stony Brook, Rommel Eng Dept, NY, Italy	State Univ of New York, Stony Brook, Rommel Eng Dept, NY, Italy	California Univ, Dept of Elect & Comput Eng, La Jolla, USA
CISE Dept, Florida Univ, Gainesville, FL, USA	Florida Univ, CISE Dept, Gainesville, USA	California Univ, Dept of Elect & Comput Eng, San Diego, USA
Comput Dept, Univ of the Highlands & Islands, Isle of Lewis, UK	Univ of the Highlands & Islands, Comput Dept, Isle of Lewis, UK	California Univ, Dept of Elect & Comput Eng, Santa Barbara, USA
Comput Eng Dept, American Univ of Sharjah, United Arab Emirates	American Univ of Sharjah, Comput Eng Dept, United Arab Emirates	Central Florida Univ, Dept of Elect & Comput Eng, Orlando, USA
Dept of Elect & Comput Eng, California Univ, Davis, USA	California Univ, Dept of Elect & Comput Eng, Davis, USA	Delft Univ of Technol, Radiat Chem Dept, Netherlands
Dept of Elect & Comput Eng, California Univ, La Jolla, CA, USA	California Univ, Dept of Elect & Comput Eng, La Jolla, USA	Dong-A Univ, Dept of Elect Eng, Busan, South Korea
Dept of Elect & Comput Eng, California Univ, San Diego, CA, USA	California Univ, Dept of Elect & Comput Eng, San Diego, USA	Dong-A Univ, Dept of Elect Eng, South Korea
Dept of Elect & Comput Eng, California Univ, Santa Barbara, CA, USA	California Univ, Dept of Elect & Comput Eng, Santa Barbara, USA	Hanoi Univ, Chic Dept, Gainesville, USA
Dept of Elect & Comput Eng, Central Florida Univ, Orlando, FL, USA	Central Florida Univ, Dept of Elect & Comput Eng, Orlando, FL, USA	Gunma Univ, Dept of Electron Eng, Japan
Dept of Elect & Comput Eng, Polytech Univ, Brooklyn, NY, USA	Polytech Univ, Dept of Elect & Comput Eng, NY, USA	Gunma Univ, Dept of Electron Eng, Kiryu, Japan
Dept of Elect & Comput Eng, Polytech Univ, New York, NY, USA	Polytech Univ, Dept of Elect & Comput Eng, New York, USA	Hanyang Univ, Appl Phys Dept, Ansan, South Korea
Dept of Elect & Comput Eng, Texas Univ, El Paso, TX, USA	Texas Univ, Dept of Elect & Comput Eng, El Paso, USA	Helsinki Univ of Technol, Theor Comput Sci Dept, Espoo, Finland
Dept of Elect & Comput Eng, Texas Univ, San Antonio, TX, USA	Texas Univ, Dept of Elect & Comput Eng, San Antonio, USA	Hong Ik Univ, Dept of Elect & Electron Eng, Seoul, South Korea
Dept of Elect & Electron Eng, Hong Ik Univ, Seoul, South Korea	Hong Ik Univ, Dept of Elect & Electron Eng, Seoul, South Korea	Hong Kong Univ of Sci & Technol, Dept of Elect & Electron Eng, Clear Water Bay, China
Dept of Elect & Electron Eng, Hong Kong Univ, China	Hong Kong Univ, Dept of Elect & Electron Eng, China	Hong Kong Univ, Dept of Sci & Technol, Dept of Elect & Electron Eng, Hongkong, China
Dept of Elect & Electron Eng, Hong Kong Univ of Sci & Technol, Clear Water Bay, China	Hong Kong Univ of Sci & Technol, Dept of Elect & Electron Eng, Clear Water Bay, China	Hong Kong Univ, Dept of Elect & Electron Eng, China
Dept of Elect & Electron Eng, Hong Kong Univ of Sci & Technol, Hongkong, China	Hong Kong Univ of Sci & Technol, Dept of Elect & Electron Eng, Hongkong, China	Inha Univ, Dept of Elect Eng, Incheon, South Korea
Dept of Elect Eng, Dong A Univ, Busan, South Korea	Dong-A Univ, Dept of Elect Eng, Busan, South Korea	Inha Univ, Dept of Elect Eng, South Korea
Dept of Elect Eng, Dong-A Univ, South Korea	Dong-A Univ, Dept of Elect Eng, South Korea	Iran Univ of Sci & Technol, Power Eng Dept, Tehran, Iran
Dept of Elect Eng, Inha Univ, Incheon, South Korea	Inha Univ, Dept of Elect Eng, Incheon, South Korea	Malaya Univ, Dept of Elect Eng, Kuala Lumpur, Malaysia
Dept of Elect Eng, Inha Univ, South Korea	Inha Univ, Dept of Elect Eng, South Korea	Malaya Univ, Dept of Elect Eng, Malaysia
Dept of Elect Eng, Malaya Univ, Kuala Lumpur, Malaysia	Malaya Univ, Dept of Elect Eng, Kuala Lumpur, Malaysia	Polytech Univ Brooklyn, Dept of Elect & Comput Eng, NY, USA
Dept of Elect Eng, Malaya Univ, Malaysia	Malaya Univ, Dept of Elect Eng, Malaysia	Polytech Univ, Dept of Electr & Comput Eng, New York, USA
Dept of Electron Eng, Gunma Univ, Japan	Gunma Univ, Dept of Electron Eng, Japan	Pusan Nat Univ, Ali Lab, Dept of Comput Sci, South Korea
Dept of Electron Eng, Gunma Univ, Kiryu, Japan	Gunma Univ, Dept of Electron Eng, Kiryu, Japan	Salank (P J) Univ, Theor Phys & Astrophys Dept, Kosice, Slovakia
Dept of Electron Eng, Shanghai Jiao Tong Univ, China	Shanghai Jiao Tong Univ, Dept of Electron Eng, China	Shanghai Jiao Tong Univ, Dept of Electron Eng, China
Dept of Electron Eng, Shanghai Jiaotong Univ, China	Shanghai Jiaotong Univ, Dept of Electron Eng, China	Shanghai Jiaotong Univ, Dept of Electron Eng, China
Power Eng Dept, Iran Univ of Sci & Technol, Tehran, Iran	Iran Univ of Sci & Technol, Power Eng Dept, Tehran, Iran	State Univ of New York, Stony Brook, Dommel Eng Dept, NY, Italy
Radiat Chem Dept, Delft Univ of Technol, Netherlands	Delft Univ of Technol, Radiat Chem Dept, Netherlands	Technische Univ München, Zentrum Math & Phys Dept, Germany
Theor Comput Sci Dept, Helsinki Univ of Technol, Espoo, Finland	Helsinki Univ of Technol, Theor Comput Sci Dept, Espoo, Finland	Texas Univ, Dept of Elect & Comput Eng, El Paso, USA
Theor Phys & Astrophys Dept, Safarik (P J) Univ, Kosice, Slovakia	Safarik (P J) Univ, Theor Phys & Astrophys Dept, Kosice, Slovakia	Texas Univ, Dept of Elect & Comput Eng, San Antonio, USA
Zentrum Math & Phys Dept, Technische Univ München, Germany	Technische Univ München, Zentrum Math & Phys Dept, Germany	Univ of the Highlands & Islands, Comput Dept, Isle of Lewis, UK

Appendix 2

Excerpts of the INSPEC Graph