Budapest University of Technology and Economics
Department of Telecommunications and Mediainformatics

Creating Heterogeneous Database Queries by Modeling Complex Semantics Encoded in Well-formed Natural Language Structures

Booklet

Zsolt T. Kardkovács

Tutors:

Dr. Gábor Magyar (BME–TMIT)
Dr. Sándor Gajdos (BME–TMIT)

Budapest, 2007

1 Introduction

The typical architecture of natural language interfaces to databases (NLIDB) consists of four layers [**?**]; of a syntactic parser, of a semantic analyser which parses syntactically analysed sentences and maps results to supporting (or partner) databases of NLIDB, of a query processor which execute queries on supporting databases, and of a ranking module which sorts results of databases queries. The core element of NLIDB is the semantic analyser that, in other words, serves as a bridge between well-formed structures and supporting databases of NLIDB.

A formal, simply and easily extensible model of the real world is an essential feature from NLIDB hence it should produce a formal representation of intentional meaning encoded in grammatically correct, well-formed inputs. This criterion is called portability in the literature [15, 11, 3], and such a model is called portable representation or portable NLIDB. The portability issues are mostly related to semantic analysis, our researches are focusing on this area.

Previous works found four approaches to achieve portable NLIDBs [11, 3]. The most of the solutions prefers dealing with language models or linguistic representations, and they are ignoring the world model encoded in supporting databases with the exception of entity–relationship model (ER-model) [9] based middle tiered NLIDB [16, 17]. Middle tiered NLIDB first transforms syntactically analysed sentences into an intermediate formal language and the intermediate language is converting into database queries. Nevertheless, ER-model based NLIDB is very sensitive for changes in database structure and for schema decompositions made in order to avoid well-known database anomalies. An ontology based middle tier arises the same problem since universal ontologies can not handle diversities of concepts on which database structures are based.

It seems reasonable to use the database world models for improving efficiency of NLIDB, specially relational databases' [10] ones. Relational data model based solutions could guarantee ontologies to be computable and as simple as possible, moreover, operations defined on data model can be applied on supporting databases without modifications. In my dissertation I am studying how to provide portability in NLIDB based on purely relation database techniques by compact representations of the real world and operations on them. My results are summarized in four theses.

Relational data model based NLIDB (RNLIDB) must comply three criteria: it should be portable, it should adaptively use yet available world models in supporting databases to create database queries by its semantic analyser, and it should model semantics defined by specific natural language structures and expressions.

The first thesis focuses on the first two criteria. The notion of portability is closely related to relational separability [8]. Therefore, I suggest a formal, mathematically decidable model instead of using the intuitive notion of portability discussed in the literature [15, 11, 3]. In order to prove separibility for wide range of relational data structures I generalise Chan–

Mendelzon theorem [8] for not dependency preserving data structures.

In the second thesis, I present a new relational model based data organisation which is portable, and which provides the necessary link between natural language representations and ontology like database structures. I show an algorithm which decomposes relational data structures into this kind of organisation. The model satifies the three criteria of RNLIDB.

I discuss how to model such a complex and ambiguous language structure like genitive phrases [5, 13] and how to transform them into SQL statements based on the introduced data structure. I prove that representation and the representation based algorithm is an improvement on processing genitive phrases hence there does not exist known procedure from the literature which is capable to transform correctly subjective or chained genitives to database queries in general.

The last thesis discusses how to handle ambiguities, relativism, heterogeneity and coverage or shift in meanings of words in databases. It states that an adequate physical organisation or indexing is capable to find the most appropriate answers in databases for some fuzzy user questions with synonymies, and with relativistic expressions. Results are language dependent as relativistic expressions are, however, they are language independent in the sense the proposed technique can be applied directly to any natural language and database.

2 Research Goals

My dissertation is aimed at creating a scalable, portable RNLIDB which re-use the world model encoded in supporting databases by construction of algorithms and models to capture semantics in natural language structures. My researches do not intent to deal with following topics and problems:

1. syntactic analsys of natural languages
2. linguistic aspects of natural language processing and understanding [1]
3. problems of information retrieval which outfit database technologies including web searches
4. study or comparision of distributed and federated databases [20] with RNLIDB

I limited my research during the design of representation techniques and models to create a widely applicable, mainly topic, language and data structure independent solution. My decisions was motivated by the intention to enhance simply my results in long terms to other application areas of information retrieval, e.g. to natural language understanding tasks.

3 Methodology

In my research I developed new models for portable RNLIDB. In order to prove features and characteristics of the model I used analytic solutions, and results in relational database

theory and set theory.

I tested the expressive power of the model by analysing text corpora and by simulations. Simulation was motivated by the fact that there is not known any NLIDB related corpus on annotated genitive phrases, and their semantics, therefore analytic methods can not be applied at all.

I proved correctness of similarity and bounded searches by analytic methods.

4 New Results

Our results and theses are discussed in three sections. The first section details a new database origanisation and its properties supporing semantic analysis by relational data model techniques. The section contains two closely related theses. The second section summarises how to model genitive phrases and how to transform them into database queries. The last section shows a new indexing technique which handles synonymy and relativism in database queries.

The most important statements, models and definitions are followed by page numbers in parenthesis which refers to pages of my dissertation where details can be found.

4.1 Foundations of Relational Model Based NLIDB

- Related publications: [J6, J3][C12, C15, C3].
- Known, independent citations: 0.
- Awards: HTE Pollák–Virág prize, 2006.

Implementations of NLIDB balance between two parameters in order to be widely applicable: the accuracy in "understanding" sentences (model and processing accuracy) and the adaptivity or portability of processing techniques in semantic analysis.

Proposed NLIDB implementations have been built on four different paradigms. Nevertheless, only one of them is dealing with information available in supporting databases which recreates ER-models of databases by reverse engineering techniques [16, 17]. The reuse of the world model of databases seems to be a reasonable and a very promising idea. Yet, it is recommended to use a more general, a quasi-ontological database model instead of database specific ER-models if one is talking about universal modeling of supporting databases since ER-model (or ontology) integration is a very hard task and an unsolved problem.

That is, we need encode the following information in our database model:

- what kinds of entities (individuals), concepts, and attributes are available in supporing databases
- what kinds of connections are defined between those entities

- what functional dependencies are defined on attributes
- which joins "make sense" on the database

The last criterion is important because e.g. the number 1038 could stand for a historical year, for a postal code, for the size of population in a village, etc., and as such joins can be applied on these attributes. However, identity of these values are only incidental, they should not join in any "reasonable" way. In the worst case, it also could lead to a data size explosion during the semantic analysis of NLIDB.

The results of my two theses are re-using large portion of database encoded information, and it also expand them by supporting database integration without any modifications on RNLIDB's data structures. The model I was created is the first successful attempt to merge fields like ontologies, databases and natural language representation techniques. My work was tested and integrated to a deep web search engine in the project called "Web of Words" under grant number NKFP-0019/2002. The project based on my solutions was awarded by the Pollák-Virág prize of Hungarian Scientific Association for Infocommunications in 2006.

Thesis I

I introduced mathematical foundations of portability for RNLIDB. I proved the introduced notion is conformed with the intuitive meaning used in literature. I proved criteria of portability for RNLIDB are decidable. I proved relational separability, and thus portability is independent from functional dependecy preserving property. I generalise the Chan–Mendelzon theorem to not dependency preserving relational data structures, and I proved this new theorem.

Portability is an intuitive notion in literature [15, 11, 3], therefore all solutions can be treated portable since neither portability nor non-portability can be proven. In the literature the following properties of portability are mentioned:

- It uses topic independent language rules for parsing.
- It provides semantic analysis on topic independent knowledge base.
- It has a general purpose lexicon.
- It deals with complex queries contain sequences of operations on database structures.
- It needs no linguistic, programming language or other information engineering skills to use the NLIDB.

Further in this paper, I use known database notions like representative instance [19], consistent database [19], independent [19] and separable [8] data structure. Their formal descriptions and their properties are widely discussed in my dissertation (pp. 13–18). I summarised the portability of semantic analysis in RNLIDB in the following way (pp. 20–21):

- Let $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ be a relational database, where $\mathbf{R}$ stands for schemas, $\mathbf{r}$ for relations, and Σ for functional and inclusion dependencies over the database, respectively. The $\mathcal{DB}$ is consistent and separable,
- the database $\mathcal{DB}' = <\mathbf{R}', \mathbf{r}', \Sigma'>$ created by any topic extension must keep consistentency and separability, and it must entail $\mathbf{R} \subseteq \mathbf{R}'$, $\mathbf{r} \subseteq \mathbf{r}'$, and $\Sigma \subseteq \Sigma'$,
- and semantic analysis should use (and parse) only the formally defined Γ language to produce database queries.

Consistency checking in relational database is a decidable [19] problem using `CHASING` procedure on representative instance [19]. One can use Chan–Mendelzon theorem [8] to prove separability of a database structure which states for any $\mathbf{R}$ schemas which preserve Σ that if there are schemas $R, S \in \mathbf{R}$ such that $R^+(\Sigma) \supseteq S$ then it is separable if and only if there exists a non-trivial $\{X \rightarrow A\} \in \Sigma$ for which X is a superkey of S, and A is an attribute of S. The theorem implies that finite number of attribute set closures is enough to decide whether $\mathbf{R}$ is separable or not. Obviously, $\mathbf{R}$ dependency preserving and independent regading Σ is also proven to be decidable in finite steps [19].

The theorem tells nothing about not dependency preserving $\mathbf{R}$ structures? I proved by examples (pp. 18) there exists a relational structure $\mathbf{R}$ which does not preserves Σ yet is separable regarding Σ, and there also exists an $\mathbf{R}$ which preserves Σ without being separable regarding it. That is, separability and dependency preserving features are independent.

For example, let $\mathbf{R} = \{R(A), S(B)\}$ be a set of schemas where R and S consist of single attributes, and let $\Sigma = \{A \rightarrow B\}$ a minimal cover of a $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ database. By definition, $\mathbf{R}$ do not preserves Σ, on the other hand, $\mathcal{DB}$ is independent and separable due to the fact that A és B are keys. As a consequence, the `CHASING` procedure can not apply the functional dependency $A \rightarrow B$, that is why the representative instance of $\mathbf{r}$ projected to $\mathbf{R}$ will always be identical to $\mathbf{r}$. That is, I proved separability does not imply and it is not implied by the dependency preserving feature.

The depedency preservation does implies separability even though $\mathbf{R}$ is lossles and independent regarding Σ. Let $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ be a database for which $\mathbf{R} = \{R(ABC), S(ABD)\}$, $\Sigma = \{A \rightarrow D, AB \rightarrow C\}$, and $\mathbf{r} = \{r(R) = \{<a, b_1, c>\}, r(S) = \{<a, b_2, d>\}\}$. $\mathbf{R}$ is independent regarding Σ since local consistency naturally implies global consistency, and it is a lossless decomposition of the universal database schema. However, $\mathbf{R}$ is not separable regarding Σ because the `CHASING` procedure creates the tuple $t[ABCD] = <a, b_1, c, d>$ for the representative instance using dependency $A \rightarrow D$, and t projected to $S(ABD)$ in not element of $r(S)$. That is, I proved dependency preservation is independent regarding separability (pp. 18–19).

As dependency preservation and separability are independent the Chan–Mendelzon theorem applicability is strongly limited. If a schema structure $\mathbf{R}$ does not preserve a set of dependencies Σ one can not decide whether $\mathbf{R}$ is separable regarding Σ or not, and

thus portability is not decidable either.

Suppose that $\mathbf{R}$ is independent regarding a set of dependencies Σ, and it does not preserve Σ. That is, there exists a dependency $X \to A \in \Sigma$ such that X and A do not appear in none of $R \in \mathbf{R}$. Since $\mathbf{R}$ is independent regarding Σ any $\mathbf{r}$ over $\mathbf{R}$ has consistent representative instance, i.e. there can not be found any tuples t_1, t_2 in the representative instance of $\mathbf{r}$ for which $t_1[X] = t_2[X]$ and $t_1[A] \neq t_2[A]$ hold. It is possible if and only if for any $\mathbf{r}$ the representative instance is not defined on attributes $X \cup \{A\}$, at least one of these attribute values must be `NULL` for any tuples in the representative instance. Indirectly, if there exists a tuple with non-`NULL` values on all attributes of $X \cup \{A\}$, and domains of attributes are large enough then it can be constructed an Armstrong relation [4] in which all functional dependencies $\Sigma' \subset \Sigma$ that preserved by $\mathbf{R}$ are hold but $X \to A$ does not – which contradicts the independecy of $\mathbf{R}$ regarding Σ.

I generalised the Chan–Mendelzon theorem (pp. 18–19) by the following statements: a database structure $\mathbf{R}$ is separable if and only if $\mathbf{R}$ is independent regarding Σ and if there exist two different $R, S \in \mathbf{R}$ schemas such that $R^+(\Sigma') \supseteq S$ – where $\Sigma' \subseteq \Sigma$ is the maximal subset of Σ that $\mathbf{R}$ preserves – there exists a non-trivial $\{X \to A\} \in \Sigma'$ for which X is the superkey of S, and A is an attribute of S.

Thesis II

I proposed an extension to relational database models which describes semantics encoded relational databases. I proved this model to be sound and complete respected to valid relationships [J3][C3]. I proposed a new data organisation for relational databases called natural databases which models natural language ambiguities and interconnections between concepts and notions used in languages [J3][C2, C3]. I showed there exist separable natural databases [C3].

It is very expensive to create universal ontologies for parsing informations hidden in natural language sentences, although there are widely available ontological knowledge in databases. Relational databases are well-known, standardized, and they have robust, simple, and decidable mathematical foundations. Nevertheless, their expressive power is far beneath ontologies. Phrasal structure of natural languages, ambiguities can not be handled by or mapped into strict database elements and limited operations. It is the reason why relational databases were not served as a source of knowledge in general for NLIDB.

RNLIDB must comply with the following three criteria: it should be portable, it should use the knowledge (semantics) encoded in supporting databases, and it should model language specific phenomena. No NLIDB describes or re-uses world models of supporting databases. The only exception were a proposition that suggested to use the ER-model of databases to capture semantics in syntactically parsed natural sentences. This approach can not be applied efficiently to integrate multiple databases, moreover, it is very sensitive for structural modifications. On the other hand, natural language phenomena were mostly modelled by ontologies and context dependent languages [3][J3] solutions do not exploit

supporting databases for this purpose [2]. The question is: is there any RNLIDB which complies the criteria?

Relational data model is "value-oriented", i.e. only the user has information on which schemas or attributes can be joined in the database, or in other words, which joins hold relevant information. Nonetheless, it is essential for NLIDB to describe all possible valid joins hence the absence of such a description implies no guarantee on creating semantically tractable, and complex queries. There is a need for model valid relationships between attributes that still missing from the literature. It is well-known that valid relationships are defined on attributes with referrential integrity and between schemas in `is-a` relationship.

That is why I introduced a new λ function (pp. 21–22) [J3][C8, C2, C7] that maps attributes sets to attribute sets with following considerations:

- for any $R \in \mathbf{R}$ schemas $\lambda(X) = X$, if $X \rightarrow R$,
- if function λ is defined on attribute sets X of $R \in \mathbf{R}$ then there exists $S \in \mathbf{R}$ such that $\lambda(X)$ is the key of S,
- $X \dashrightarrow \lambda(X)$ holds for any X attributesets if $\lambda(X)$ is defined.

I also introduced the notion Ξ in order to capture binary `is-a` relationships (pp. 21). The $\Xi(R, S)$ is said to be hold for $R, S \in \mathbf{R}$ schemas in a $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ database if and only if there exist proper subsets X and Y attributesets of R and S respectively, such that $\Sigma \models \{X \rightarrow R, Y \rightarrow S, X \dashrightarrow Y\}$. I called $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ as a semantic model of a database in which Ξ is defined for all pairs of schemas, and for all attributes A there exists an attributeset X of a schema R such that $A \in X$ and $\lambda(X)$ is defined (pp. 21–22).

I introduced the binary relation $\varepsilon(X, Y)$ to capture notion of valid relationships. The $\varepsilon(X, Y)$ is said to be true in a database $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ if and only if $\Sigma \models \{X \dashrightarrow Y, Y \rightarrow S\}$ where $X \dashrightarrow Y$ stands for an inclusion dependency. Obviously, relationship determined by Ξ or λ are valid. That is, semantic models of a database are sound respected to valid relationships. Semantics models of databases are complete regarding valid relationships as well because whenever attribute domains are large enough there can be constructed an Armstrong relation which contradicts to dependencies $X \dashrightarrow Y$ or $Y \rightarrow S$ if they are not entailed by Σ [7, 12]. As a consequence, semantic models of databases are sound and complete respected to valid relationships, all $\varepsilon(X, Y)$ are implied by Σ (pp. 22–23).

I showed (pp. 24–26) specific semantic models of databases called natural databases capture several natural language phenomena like connectivity between language concepts, ambiguities, and meaning based context idenficiation by introducing the notion of natural keys [J3][C8, C7]. Since natural keys consist of single attributes (names) in natural languages, I assumed all keys in natural databases are simple, and therefore all values of reference functions are single attributes. A basic property of natural keys is it can be ambiguous as they are in natural communications. Disambiguation is guaranteed only

by means of contexts: meaning of words and the grammatical structure of sentences [J6][C12, C15, C3].

I defined natural databases in equivalent ways (pp. 25). I call semantic models of databases natural if all keys are natural keys. We determine the same set of semantic models of databases if the function λ has a values set consist only of natural keys. And finally, it is also equivalent to the proposition that states only those semantic models of databases are natural in which for all $\varepsilon(X, Y)$ relationships Y is a natural key.

A natural database are not necessarily separables regarding a set of dependencies Σ, and so portability is in question as well. Portability requires RNLIDB to be as decomposed as possible in order to guarantee additivity of database structures, and by avoiding further decompositions. It is easy to see that all attributes can be extracted from schemas but keys. As a consequence I suggested a new data organisation which consists of primary schemas with a single attribute their natural key, and secondary schemas serve as connect tables between primary schemas. This idea is analoguous to ontologies where logical concepts and roles have the same function. The main difference is secondary schemas not necessarily representing binary relations.

According to these considerations I introduced the notion of normalised natural database (NNDB) $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ [C3] (pp. 32–34):

1. Any schema $R \in \mathbf{R}$ which contains a natural keys has no other attribute. In this case R is called primary schema.

2. If a schema has no natural keys, it is called secondary schema, and it has at least two attributes, and all of these refers to keys of primary schemas.

3. For all distinct $R, S \in \mathbf{R}$ schemas $R \cap S = \emptyset$.

4. There are no secondary schemas $R, S \in \mathbf{R}$ for which $R^+(\Sigma) \supseteq S$ is true.

5. All secondary schema is in Boyce-Codd normal form [10, 14].

6. For all inclusion dependency $X \dashrightarrow Y$ Y is a natural key.

NNDB is separable (pp. 34). Firstly, it is because no attributes are the same and therefore the statement is implied by the Chan–Mendelzon theorem. Nevertheless, this feature is required by other considerations (see Thesis 3). Secondly, NNDB is separable because fourth requirement meets the prerequisites of generalised Chan–Mendelzon theorem. If attributes may appear in multiple schemas the fourth requirement intents to schemas be as minimal as possible to gather only related attributes in a single schema. It is worth to mention that NNDB is not dependency preserving in general still it is not an important feature of a knowledge base.

Since NNDB is separable an NNDB based RNLIDB which translates a formal Γ grammar into database queries is portable by definition if all extension of the database maintains the

NNDB properties. That is, I proved the existence of RNLIDB. I also proposed a universal algorithm which transforms relational database structure into NNDB (pp. 26–31).

4.2 Algorithmic Proccesing of Marked Genitive Phrases

- Related publications: [J3][C8, C2, C7, C3].
- Known, independent citations: 0.
- Awards: Publication [C2] was ranked as top 5 on NLIDB 2005 conference.

Genitive phrases can express wide range of semantic connections and they could have various meanining (see Table 4.1 and [6]). Moreover, possessed and possessor roles may vary in different contexts (e.g. author's book, book of an author), so algorithmic procession of such phrases is a hard task. See 4.2 for illustrative examples.

Types of genitives	Examples (Hungarian and English)
origination	Moszkva küldötte (men of Rome)
materials	– (ring of gold)
partitive relations	a tanszék vezetője (head of department)
quantitive descriptions	húsnak kilója (pound of beer)
relationship	Péter felesége (Pam's address)
ownership	Sára sapkája (John's coat)
subjectivity	Verdi operája (dramas of Shakespeare)
objectivity	Budapest látképe (portray of Elisabeth II)
function or purpose	dolgozók iskolája (school of girls)
linkage or chaining	Ábel apjának barátja (name of Tom's wife)

Table 4.1: *Different types of genitive phrases*

Universal processing of genitive phrases – both language dependent and independent solutions – is missing from the literature. Its absence is responsible for higher error rates of indexers and classifiers. That is why questions like "When did children of Juan Carlos I born?" can not be answered by web search engines and question answering tools. My dissertation shows a new NNDB based model and algorithm developed by me and widely published at both on Hungarian and international conferences and other forums. The main differences between my propositions and others' work is seen at Table 4.3.

The first detailed publication on the algorithm was selected top 5 on the solely NLIDB related conference in 2005.

Genitive expression	Equivalent SQL statement
Bizet's Carmen	`SELECT title FROM operas WHERE composer = 'Bizet' AND title = 'Carmen'`
dramas of Shakespeare	`SELECT title FROM dramas WHERE author = 'Shakespeare'`
Edith's address	`SELECT address FROM addresses WHERE name = 'Edith'`
books' roles	`SELECT role FROM roles WHERE play IN ( SELECT title FROM books )`
heads of companies	`SELECT head FROM companies`
name of Petőfi's mother	`SELECT name FROM persons WHERE name IN ( SELECT mother FROM familytrees WHERE descendant = 'Petőfi' )`

Table 4.2: *Genitive phrases and equivalent SQL statements*

Implementations	origination	materialism	partitives	quantitives	relationship	ownership	subjectivity	objectivity	purposes	chaining
Practice	n	n	y	n	y	n	n	n	n	n
START	y	n	n	n	y	n	n	n	n	n
SQ-HAL	n	n	n	n	y	n	n	n	n	n
NL for Cindi	n	n	n	n	y	n	n	n	n	n
Masque/SQL	n/a	n	y	n	y	n	n	n	n	n
NChiql	n	y	y	n	y	n/a	n	n/a	n	n
KID	n	n	n	n/a	y	n	n	n	n	n
(V)ISA	y	n	y	n	y	y	y	y	n	y

Table 4.3: *Genitive Types Processed by Known Implementations*

Thesis III

I developed a new NNDB based mathematical model called (V)ISA-model to capture semantics of marked genitive phrases. I proved this model to be the first which is capable to model genitive phrases expressing objectivities and to deal with compound (or chained) genitive phrases. I proposed an algorithm which transform genitive phrases into SQL statements. I proved this algorithm to be sound respected to the (V)ISA-model. I also tested the accuracy of the model.

The (V)ISA-model and (V)ISA-algorithm which is based on the model are language

independent in the sense they do not depend on the linguistic context they are applied on, nevertheless, they contain a language dependent relationship that treats differences of genitive phrases in distinct languages. This relations is not my result (see [6, 5, 21]), however, the extension of the relationship to meaningless language dependent expressions is a new proposition.

Genitive phrases in sentences are transformed into embedded SQL statements in general. Embedded queries are far beyond the known NLIDB capabilities. One of the reasons is there is not known any universal model which captures hidden semantics in genitive phrases. My model is the first attempt to create one to handle such an important language structures like genitives are.

Let us introduce symbols $\mathcal{A}$ and $\mathcal{I}$ besides $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma, \lambda, \Xi>$ for the set of attributes, and for the set of individuals (entities) described in $\mathcal{DB}$, respectively, for a better understanding. Let $\alpha{\twoheadrightarrow}\beta$ stand for a genitive phrase in which α is the possessor and β is the possessed part.

Entities with possessed roles in genitives are mostly individuals, set of individuals, or primary schemas. Primary schemas can be treated as set of individuals defined by the relations over primary schemas. As a consequence I modelled genitive phrases $\alpha{\twoheadrightarrow}\beta$ as α is a set of individuals while β could refer to either schemas, attributes or individuals (pp. 40–42). Nota Bene, some genitive phrases are valid in a given language and some others are not [5, 18, 22], therefore a new relationship Π is introduced to identify valid possessives. $\Pi(\alpha, \beta)$ is said to be valid or true if and only if there exist a valid natural language sentence of a given language in which $\alpha{\twoheadrightarrow}\beta$ have meaning (pp. 42–43).

But what do genitive phrases represented in databases mean or result?

Let $\|A\| = \{B \mid B \in \mathcal{A} \ \wedge \ \lambda(B) = \lambda(A)\}$ stand for the equivalence class of an attribute $A \in \mathcal{A}$ in NNDB. Since values of λ are natural keys it really defines an equivalence class. Let $\Psi : 2^{\mathcal{A}} \to 2^{\mathbf{R}}$ be a mapping which determines the maximal set of schemas in which elements of attribute sets are contained. That is,

$$\Psi(X \in 2^{\mathcal{A}}) = \{R \mid \exists A \in X \wedge A \in R\}.$$

Let us introduce $\psi : \mathcal{I} \to 2^{\mathcal{A}}$ such that

$$\psi(I \in \mathcal{I}) = \{A \mid I \in \mathrm{DOM}(A)\}.$$

if β is a possessed part of a genitive phrase then γ stand for the following expression depending on β:

$$\gamma = \begin{cases} \psi(\beta) & \text{ha } \beta \subseteq \mathcal{I} \\ \|\kappa(\beta)\| & \text{ha } \beta \in \mathbf{R} \\ \|\beta\| & \text{ha } \beta \in \mathcal{A} \end{cases}$$

Due to the introduced notions I modelled the genitive phrase $\alpha{\twoheadrightarrow}\beta$ in the following way:if there exist $A, B \in \mathcal{A}$ for which $A \in \psi(\alpha)$, $B \in \gamma$ such that $R \in \Psi(\psi(\alpha)) \cap \Psi(\gamma)$, $\{A, B\} \in R$,

and $\Pi(A, B)$ are true then $\alpha{\twoheadrightarrow}\beta = \mathrm{DOM}(B) \cap \mathcal{I}$ assuming α is a single individual. If α represents a set of individuals then

$$\alpha{\twoheadrightarrow}\beta = \bigcup_{i \in \alpha} i{\twoheadrightarrow}\beta.$$

On the basis (V)ISA-model I developed the (V)ISA-algorithm which translates $\alpha{\twoheadrightarrow}\beta$ genitive phrases into SQL statements (see Algorithm 1)[J3][C8, C2, C7]. The function κ in the pseudo-code results the natural key of a primary schema (pp. 44–46). It is seen in pseudo-code that definitions of (V)ISA-model is calculated by (V)ISA-algorithm step-by-step, therefore the algorithm is obviously sound respected to the model. I gave illustrative examples on (V)ISA-algorithm in action, and I proved the algorithm produces SQL statements for phrases equivalent to those in Table 4.2.

Using NNDB an average complexity is around $O(n(\log|\mathcal{I}| + \log|\mathcal{A}|))$ where n is the number of search requests (pp. 44). I showed that individuals and attribute values can be search by logarithmic complexity with a proper representation and indexing. Such a representation is discussed on pages 35–37. If we assume that results for the most of queries are limited in number then we can declare (V)ISA-algorithm has a logarithmic complexity. In some cases, processing time depends on possessed part of genitives is a schema, an attribute or an individual, and on the size of the resulting set. As a consequence, algorithm has $O(|\mathcal{I}|)$ worst case complexity.

The effectiveness of the algorithm was implicitly tested in the project called "Web of Words" (NKFP-0019/2002). Firstly, because the accuracy of RNLIDB depends on other components of the systems, e.g. on syntactic parser. Secondly, there are not available any databases or corpora for genitives in none of the languages, there are not known other genitive model or learning method (V)ISA-model can be compared with. That is, I made comparisons on expressive powers only (see Table 4.3 for details). Table 4.3 contains value y if there can be found many witnesses for implementations dealing with a given genitive type. I found (V)ISA-algorithm unique to support genitive phrases expressing objectivity and chained possessives.

I tested the algorithm by typical and "malignant" user questions as well. I found (pp. 47–49) genitive phrases represented in databases are modelled with precisions 78.87% for all questions, and with 91.83% for typical queries. The test indicates that metaphoric and complex expressions can not be captured by (V)ISA-model (e.g. *Mecca of movies*, *winner of a game*). Further details are seen in the dissertation.

4.3 Similarity Search in Databases

- Related publications: [J1, J2, J4, J5][C4, C6, C1, C13].
- Known, independent citations: 7.
- Awards: algorithm which incorporated similarity search was awarded by runner-

Algorithm 1: The (V)ISA algorithm

```
1  Given an NNDB DB = <R, r, Σ, λ, Ξ> with individuals I and attributes A .
2  function VISA( α, β) returns SQL;
3  begin
4      if  β ∈ A then  γ := ‖β‖;
5      else if  β ∈ R then  γ := ‖κ(β)‖;
6      else  γ := ψ(β);
7      forall  R ∈ Ψ(ψ(α)) ∩ Ψ(γ) do
8          forall  A ∈ ψ(α) ∩ R do
9              forall  B ∈ γ ∩ R do
10                 if  Π(A, B) then
11                     post := ( β ∈ A ) ? " and 'B = β'" : "'";
12                     head := 'select B from R ';
13                     if  Set( α) then
14                         return head + 'where A = α' + post;
15                     end
16                     else
17                         δ := ( α = γ↠ε ) ? VISA( γ, ε) : α;
18                         if  α ∈ R then
19                             return
                               head + 'where A in ( select κ(α) from α )' + post;
20                         end
21                         else
22                             return head + 'where A in (δ)' + post;
23                         end
24                     end
25                 end
26             end
27         end
28     end
29     return '';
30 end
```

up awards in categories precision and creatitivity by Association for Computing Machinery (ACM) on the Knowledge Discovery and Data Mining Cup (KDDCup) 2005

A felhasználók és az adatbázist építő-szervező szakértők közötti számottevő fogalomhasználatbeli eltérés lehet, különös tekintettel a fogalmak által jelölni szándékozott jelentéstartalomra. Természetesen, a jelenség nem csak lekérdező és adatbázis, hanem akár hasonló tematikájú, tartalmú adatbázisok közötti viszonyokra is fenn állhat. Különösen problémás a helyzet az olyan elnagyolt jelentésű fogalmakkal és nyelvi szerkezetekkel, mint amilyen pl. az olcsó, magyar, nagy vagy a közel lenni valamihez. Az eltérő, bizonytalan fogalomhasználatból azonban adódik, hogy a szakszerűen létrehozott és kezelt adatbázistartalom nem feltétlenül egyezik meg a felhasználói kérdésben rögzített fogalmakkal. Adatbázisokban csak és kizárólag egzakt kereséseket lehet végrehajtani – tipikusan néhány keresési kulcs alapján, ennek megfelelően egy a jelentéstartalmi különbség kezeléséhez szükség van a közelítő keresés támogatására adatbázisokban.

Az NLIDB természetes kiterjesztése, illetve a felhasználói viselkedés általános támogatása lehet egy olyan rendszer, amely a felhasználótól nem igényli még az adatbázisban kereshető, megfelelő szakkifejezések, névelemek ismeretét sem. Kutatásaim arra irányultak, hogy ezt a fajta kiterjesztést általánosan, egy új indexelési eljárással valósítsam meg. Az általam javasolt megoldás, a természetes nyelvi feldolgozás strukturált adatszerkezetek, azon belül is, NLIDB gerincét képező adatbázisok általi támogatása, egy teljesen új irányt képvisel.

Thesis IV

Eldönthető, többkulcsos, intenzionális attribútumokra is kiterjeszthető új, előrendezési relación alapuló indexelési eljárást alkottam a nyelvi jelentéstartalmi átfedések és fokozások számítógépes modellálására. Megmutattam, hogy a modell irányított körmentes gráfként reprezentálható. Igazoltam, hogy a gráf alapú indexelésre épített keresési eljárás az attribútumok sorrendjére invariáns, illetve támogatja az indexelt attribútumok részhalmazának, azaz részkulcsok keresését is $O(k \log_k n)$ *költséggel* $k > 1$ *esetén, ahol* k *a gráf átlagos fokszáma és* n *a gráf pontjainak a száma. A modellt kiterjesztettem hasonló, közelítő és korlátos értékek több kulcs alapján való keresésére. A modellhez konstruktív módon megalkottam egy új keresési eljárást, amelyről igazoltam, hogy helyes és teljes a modellre nézve.*

Több attribútum szerinti hasonlóság keresés támogatására az adatbázisok esetében rangsoroló (ranking) eljárásokat szoktak alkalmazni, ezek azonban elsősorban a találati pontosságot, az adatbázisban tárolt és a lekérdezésben rejlő kifejezéseket közelíti egymáshoz. Sokkal általánosabb megoldást kapunk, ha az adatbázisban tárolt értékek között egy általános rendezési relációt definiálunk. Az alkalmazott indexelési eljárás újszerűsége az, hogy az adatbázisok elméletében kevéssé kutatott ún. előrendezési (pre-order)

reláció tulajdonságaira építi fel a kereséshez használt indexfát, amely reláció jóval általánosabb a nemzetközi szakirodalomban használt rendezési relációknál (51–52. oldal). A modellnek előnye társaihoz képest, hogy képes jelentésbeli eltolódásokat, fokozódások, résztulajdonságok kezelésére éppúgy, mint származtatott, intenzionális attribútumok rendezésére – eddig egyedülálló módon. A modellről elmondható – bár ez nem része a tézisnek –, hogy az Bernays-Schönfinkel-Ramsey osztályba, ennek megfelelően az eldönthető modellek közé tartozik [J2].

Legyen adott egy $\mathcal{DB} = <\mathbf{R}, \mathbf{r}, \Sigma>$ adatbázis. Jelölje $ri(\Omega)$ $\mathcal{DB}$ adatbázis reprezentatív példányát, ahol Ω a $\mathcal{DB}$ adatbázis univerzális sémájának felel meg. Az Ω sémáról feltételezzük, hogy 0NF, azaz attribútumértékei értékhalmazok is lehetnek, míg a `NULL` kitöltéseket mindenhol üres halmaznak ($\emptyset$) tekintjük.

Legyen t_1, t_2 egy–egy ennese az Ω sémára illeszkedő $ri(\Omega)$ relációnak, továbbá legyen ϱ egy olyan bináris, tranzitív és reflexív (előrendezési) reláció, amelyre

$$\varrho : \mathrm{DOM}(A) \times \mathrm{DOM}(A) \rightarrow \{\top, \bot\},$$

ahol $\top$ a logikai igaz, míg $\bot$ a logikai hamis szimbólumot jelöli, és $A \in \Omega$. Bevezettem, hogy t_1 helyettesíthető t_2-vel valamely A attribútumot és a ϱ relációt tekintve, ha $\varrho(t_1[A], t_2[A]) = \top$, és ezt a tényt $t_1 \preccurlyeq_A^{\varrho} t_2$-vel jelöltem.

A helyettesíthetőséget attribútumhalmazokra is kiterjesztettem (52–53. oldal), ezt elneveztem fedésnek. Legyen t_1, t_2 egy–egy ennese valamely Ω sémára illeszkedő $r_i(\Omega)$ relációnak, továbbá legyen $\Phi = \{\varrho_1, \varrho_2, \ldots, \varrho_n\}$ egy olyan bináris, tranzitív és reflexív relációkból halmaz, amelyre

$$\varrho_i : \mathrm{DOM}(A_i) \times \mathrm{DOM}(A_i) \rightarrow \{\top, \bot\},$$

bármely i indexű ϱ_i relációra és különböző $A_i \in \Omega$ ($A_i \neq A_j$, ha $i \neq j$) attribútumra. Azt mondjuk, hogy t_1 fedhető t_2-vel valamely $R \subseteq \Omega$ attribútumhalmazra és Φ relációhalmazt tekintve, ha

$$\forall A_i \quad A_i \in R \Rightarrow t_1 \preccurlyeq_{A_i}^{\varrho_i} t_2,$$

, ahol ϱ_i az A_i attribútumnak megfelelő reláció. Ezt a tényt $t_1 \sqsubseteq_R^{\Phi} t_2$-vel jelöljük.

Az előrendezési reláció gráfként reprezentáltam (53. oldal). Legyen $\mathcal{G}_R^{\Phi} = <\mathbf{V}, \mathbf{E}>$ egy irányított gráf, amelyben $\mathbf{V}$ elemei attribútumértékek rendezett halmazai, míg $<v_i, v_j> \in \mathbf{E}$ akkor és csak akkor, ha $v_i, v_j \in \mathbf{V}$, $v_i \neq v_j$ és $v_i \sqsubseteq_R^{\Phi} v_j$. Az ilyen tulajdonságú $\mathcal{G}_R^{\Phi}$ gráfot teljes Φ/R–katalógusnak neveztem el. Mivel az előrendezés tranzitív, ezért értelemszerűen a teljes Φ/R–katalógus erősen összefüggő részei klikkeket alkotnak.[J2][C6]. A tranzitivitás redundáns tárolást jelent a gráf tárolása esetében, ezért érdemes a kikövetkeztethető élekkel redukálni a gráfot.

Legyen $\mathcal{G}_R^{\Phi} = <\mathbf{V}, \mathbf{E}>$ egy teljes katalógus, amelyből a redukált $\mathcal{G}'^{\Phi}_R = <\mathbf{V}', \mathbf{E}'>$ katalógust (röviden katalógust) úgy kapjuk meg, hogy $\mathbf{V}'$ elemei $\mathcal{G}_R^{\Phi}$ klikkjeinek felelnek

meg, és $v'_i, v'_j \in \mathbf{E}'$ akkor és csak akkor, ha $v'_i \neq v'_j$ és

$$\exists v_i \exists v_j \forall v_k \qquad v_i \in C_i \ \wedge \ v_j \in C_j \ \wedge \ v_k \in C_k \ \wedge \ C_i \neq C_j \ \wedge \ C_i \neq C_k \ \wedge \ C_j \neq C_k \ \wedge$$
$$\wedge <v_i, v_j> \in \mathbf{E} \Rightarrow <v_i, v_k> \notin \mathbf{E} \vee <v_k, v_j> \notin \mathbf{E},$$

ahol v'_i, v'_j a $C_i, C_j \subseteq \mathbf{V}$ klikkeknek megfelelő csomópontok és $C_k \subseteq \mathbf{V}$ tetszőleges klikk (53–54. oldal). Az így kapott gráf irányított körmentes gráf (DAG - Directed Acyclic Graph).

Könnyen belátható, hogy a katalógus támogatja a részkulcs alapú keresést is, hiszen a rendezési reláció az attribútumok összességére, így azok részhalmazaira is érvényes. A gráf segítségével szemléltethetjük a keresés legfontosabb lépéseit. Tegyük fel, hogy $R' \subset R$ nem üres attribútumhalmaz mint részkulcs alapján szeretnénk keresni. Ha a tetszőleges v_j csomópontban vagyunk, akkor a részkulcs alapú kereséshez minden olyan v_k csomópontot meg kell vizsgálnunk, amelyhez létezik $<v_j, v_k> \in \mathbf{E}$ irányított él, és amelyre a $v_j \sqsubseteq^{\Phi}_{=R'} v_k$. Legrosszabb esetben valamennyi v_k gyökerű részfát be kell járnunk, ami akár a teljes gráfot is jelentheti. Átlagos esetben, ha a gráf átlagos fokszáma k, akkor $k/2$ elágazást járunk be minden csomópontban. Mivel az elágazások számával a gráfban található irányított utak átlagos hossza logaritmikusan csökken – kiegyensúlyozott gráfot feltételezve –, ennek alapján a részkulcs alapú keresés átlagos komplexitására $O(k \log_k n)$ adódik, amennyiben $k > 1$. Ha $k \leq 1$, akkor $O(n)$ átlagos költsége van a részkulcskeresésnek, hiszen ebben az esetben legfeljebb lineáris rendezésről beszélhetünk (54–56. oldal).

A relációs adatbázisok érték szerint kapcsolnak össze különböző valós világbeli megfigyeléshez, entitáshoz tartozó rekordokat. Ennek megfelelően a relációs adatbázisok megkövetelik a felhasználóktól, hogy a világbeli elemeket pontosan úgy nevezzék, ahogyan az adatbázisban tárolva, rögzítve van. Az NLIDB világában jellemzően több adatbázist kapcsolunk össze, így nem várható el sem az adatbázist létrehozóktól, sem a felhasználóktól, hogy minden körülmények között egyértelmű, félreérthetetlen fogalmakat használjanak, hiszen ebben az esetben az embert rendelnénk a gép szolgálatába, ami nem mondható kívánatosnak.

Példának okáért, ha valaki tengerparti nyaralást tervez, és egy olyan keresőkifejezést szeretne megfogalmazni, hogy a szállodája legyen közel a tengerhez, ne legyen túl drága, legyen elérhető számos szórakozási lehetőség, akkor bizony egy utazási iroda jól felkészített adatbázisa sem tudna megfelelő ajánlatot adni az ügyfél számára. Akkor sem, ha a közelit mondjuk 200 méternél kisebbnek, a nem túl drágát kb. 100.000 forintos árnak, a számos fogalmát pedig legalább négynek értelmezzük. Az adott kérdésre könnyen előfordulhat, hogy nem fogunk választ kapni egy lekérdezéssel, hiszen pontosan ilyen ajánlat nem feltétlenül van az adatbázisban; valahogyan modellálni kellene, hogy mi lehet a rekord- vagy adathasonlóság.

Az adathasonlóságot kizárólag speciális, adott attribútum szemantikáját teljesen kihasználó rangsoroló algoritmusok segítségével szokás osztályozni. A rangsoroló algoritmusok azonban nem alkalmazhatóak hatékonyan többkulcsos keresés, illetve

intenzionális és halmazértékű esetében sem. Az adathasonlóságnak ezért egy jóval általánosabb modelljét [J2] alkottam meg, a $\mathcal{G}_R^\Phi$-katalógusok felhasználásával (56–61. oldal).

Legyen $\mathcal{G}_R^\Phi = <\mathbf{V}, \mathbf{E}>$ egy redukált katalógus és K egy keresőkifejezés. Vezessük be az alábbi jelöléseket és fogalmakat:

- Jelölje $\min(K)$ azon csomópontok maximális halmazát – ezt K keresőkifejezés alsó korlátjának fogjuk nevezni –, amelynek elemei az alábbi tulajdonsággal bírnak:

  $$\min(K) = \{v \mid v \in \mathbf{V} \wedge \exists v_i \quad v_i \in \mathbf{V} \wedge <v, v_i> \in \mathbf{E} \wedge \neg v_i \sqsubseteq_R^\Phi K \wedge v \sqsubseteq_R^\Phi K \sqsubseteq_R^\Phi v_i\},$$

  ahol $\neg$ a logikai tagadás jele.

- Hasonlóan jelölje

  $$\max(K) = \{v \mid v \in \mathbf{V} \wedge \exists v_i \quad v_i \in \mathbf{V} \wedge <v_i, v> \in \mathbf{E} \wedge \neg K \sqsubseteq_R^\Phi v_i \wedge v_i \sqsubseteq_R^\Phi K \sqsubseteq_R^\Phi v\}$$

  a gráf adott tulajdonságú csomópontjainak maximális halmazát, amelyet K keresőkifejezés felső korlátjának fogunk nevezni.

- Jelölje végül $\mathrm{sim}(K)$ azon csomópontok maximális halmazát, ezeket K-hoz hasonló adatelemeknek nevezzük, amelyekre ha $\max(K) \cap \min(K) = \emptyset$, akkor

  $$\begin{aligned}\mathrm{sim}\ (K) = \quad & \{v \mid v \in \mathbf{V} \wedge \exists v_i \quad (v_i \in \max(K) \cup \min(K)) \vee \\ & (v_i \in \max(K) \ \wedge \ <v, v_i> \in \mathbf{E}) \ \vee \ (v_i \in \min(K) \ \wedge \ <v_i, v> \in \mathbf{E}) \ \},\end{aligned}$$

  egyébként pedig $v_K = \max(K) \cap \min(K)$ $(v_k \neq \emptyset)$ esetén

  $$\begin{aligned}\mathrm{sim}\ (K) = \quad & \{ \ v \mid v \in \mathbf{V} \wedge \exists v_i \quad ((<v_i, v_K> \in \mathbf{E} \ \wedge \ <v_i, v> \in \mathbf{E}) \vee \\ & \vee \ (<v_K, v_i> \in \mathbf{E}) \ \wedge \ <v, v_i> \in \mathbf{E}) \ \}.\end{aligned}$$

Az egyszerű keresési algoritmus [J2][C4] (55. oldal) működési elve szerint – a gyökérből indulva – egy olyan elemet keres, amelyet a keresőkifejezés fed. Ha ilyen elem nincs, akkor értelemszerűen a gráf csomópontjai által reprezentált ennesek egyike sem felel meg a keresési feltételnek. Ha mégis akadna ilyen, akkor a tranzitív tulajdonság felhasználásával az új kezdőpontnak a talált csomópontot veszi. Ha az adott csomópont fedi a keresőkifejezést, akkor a csomópont által reprezentált enneseket kerestük. Ha nem, akkor rekurzívan folytatjuk a keresést a talált csomópontból mint új gyökérből kiindulva.

A hasonló adatok feltérképezésének algoritmusát az általam alkotott hasonlósági kereső eljárás [J2][C1, C6] írja le (58. oldal), amely persze részben felhasználja az egyszerű keresésnél már alkalmazott `Trunc` eljárást. Az algoritmus megkeresi a gráfban azt a helyet, ahol a keresőkifejezésnek megfelelő csomópont a gráfban lennie kellene. Ha van olyan csomópont, amely a keresőkifejezésnek megfeleltethető, akkor az adott csomóponton kívül a hozzá kapcsolódó elemeket, ha ilyen csomópont nincs, akkor pedig a neki megfelelő csomópontot virtuálisan, a gráf szemantikájának megfelelően „beszúrva” a gráfba a virtuális csomópont környezetét adja vissza.

Disszertációban bizonyítottam, hogy ha a $\mathcal{G}_R^\Phi = <\mathbf{V}, \mathbf{E}>$ egy redukált katalógus, akkor a $v_i, v_j \in \mathbf{V}$ csomópontjaira $v_i \sqsubseteq_R^\Phi v_j$ akkor és csak akkor áll fenn, ha van a gráfban egy irányított út v_i és v_j között (54. oldal). Kimondtam, ha létezik irányított $v_i, v_j \in \mathbf{V}$ csomópontjai között úgy, hogy $v_i \sqsubseteq_R^\Phi v_j$, akkor a gráfban levő v_i és v_j között található összes irányított út tartalmazza az összes olyan $v_k \in \mathbf{V}$ csomópontot, amelyre $v_i \sqsubseteq_R^\Phi v_k \sqsubseteq_R^\Phi v_j$ (56. oldal). Az állítás bizonyítását a disszertációban megtalálható. E két állítás segítségével igazoltam, hogy a hasonlósági keresés algoritmusa helyes, azaz megfelelő K keresőkifejezésre pontosan az adathasonlóság definíciója szerinti $\mathrm{sim}(K)$ elemhalmazokat adja eredményül (57–59. oldal).

Az eljárás átlagos költségét nagyon nehéz meghatározni, hiszen ez nagyban függ a gráf és a keresőkifejezés sajátosságaitól. Annyi azonban elmondható, hogy legrosszabb esetben a gráf összes élét be kell járni; ekkor $O(|E|)$ komplexitásról beszélhetünk (59–60. oldal). Ha a keresendő kifejezés az adatbázisban megtalálható, akkor annak átlagos keresési költsége az egyszerű kereséssel azonos.

A KDDCup 2005 verseny kiadott adathalmazain tesztekkel igazoltam, hogy tudásszegény környezetben, dokumentumok hierarchikus kategóriafába való rendezése esetében ha rögzítjük az osztályozó beállítási paramétereit, akkor a katalógusok alkalmazása 12%-os javulást eredményez (64–68. oldal). Az alkalmazott megoldásban a katalógusok előrendezési relációja az egyes kategóriákra jellemző szavak ún. tf-idf (term frequency, inverse document frequency) értékeire vonatkozó rendezési reláció volt [J4, J5][C13].

5 Eredmények alkalmazhatósága

A 4.1 és 4.2 szakaszokban bemutatott eredményeket a NKFP–0019/2002 jelű, sikeresen lezárult Szavak hálójában projekt keretében értem el. Az eredmények javítása, bővítése jelenleg is folyamatban van, hiszen két Gazdasági Versenyképességi Operatív Program, egy Jedlik Ányos pályázat és a Mobil Innovációs Központ keretei között folytatott kutatásai tevékenységemben is döntően ezeket az eredményeket használom fel olyan területeken, mint pl.:

- gazdasági tartalmú hírek összetartozó szálainak felderítése,
- webkereső motorok indexelési technikájának javítása,
- gazdasági tartalmú hírek tényadatainak automatikus kinyerése, formalizálása,
- távközlési hívásbejegyzések problémacentrikus, tematikus osztályozása,
- mobilszolgáltatások dinamikus, futási idejű összekötettése,
- hely– és környezetfüggő mobilszolgáltatások biztosítása.

Mindegyik problémakör kulcseleme az ún. kontextusazonosítás és -felismerés, ahol az NNDB-re épülő, de nem feltétlenül természetes nyelvi, hanem pl. lokalitás, erőforrás, téma, kategória fogalmakat központba állító architektúrák sikeresen vizsgáznak.

Az algoritmusokat Java programnyelven, míg a relációs adatbázishoz köthető megoldásokat, teszteléseket Oracle adatbázis-kezelővel valósítottam meg. A megoldás időközben integrálódott egy ún. vizuális tezauruszon alapuló metakereső rendszerrel is. Jelenleg a teljes implementáció nyilvánosan még nem elérhető.

A 4.3 szakaszban bemutatott megoldást az ACM éves rendes adatbányászati rendezvényén, a KDDCup versenyen 2005–ben második helyezést elért tudásszegény szövegkategorizáló [J4, J5][C13] egyik fontos elemét képezte. Az algoritmus elnyerte a kreatív ötlet kategóriában is a II. díjat.

Bibliography

[1] Allen, J. *Natural Language Understanding (2^{nd} edition)*. The Benjaming/Cummings Publishing, Redwood City, CA, US, 1995.

[2] Amble, T. BusTUC: a natural language bus route oracle. In *Proceedings of the 6^{th} conference on Applied natural language processing* (San Francisco, CA, USA, 2000), Morgan Kaufmann Publishers Inc., pp. 1–6.

[3] Androutsopoulos, I., Ritchie, G. D., and Thanisch, P. Natural language interfaces to databases – an introduction. *Journal of Natural Language Engineering 1*, 1 (July 1995), 29–81.

[4] Armstrong, W. W. Dependency structures of data base relationships. In *IFIP Congress* (1974), J. L. Rosenfeld, Ed., North-Holland, pp. 580–583.

[5] Barker, C. *Possessive Descriptions*. PhD thesis, University of Carolina, Santa Cruz, Department of Linguistics, 1995.

[6] Barker, C., and Dowty, D. R. Non-verbal thematic proto-roles. In *Proc. of NELS 23 Conference* (Amherst, Massachusetts, 1992), North-Eastern Linguistics Conferences, GLSA Publications, pp. 49–62.

[7] Beeri, C., Dowd, M., Fagin, R., and Statman, R. On the structure of armstrong relations for functional dependencies. *Journal of ACM 31*, 1 (1984), 30–46.

[8] Chan, E. P. F., and Mendelzon, A. O. Independent and separable database schemes. In *PODS'83: Proceeding of the 2^{nd} ACM SIGACT-SIGMOD symposium on Principles of database systems* (New York, NY, USA, 1983), ACM Press, pp. 288–296.

[9] Chen, P. P.-S. The entity-relationship model – toward a unified view of data. *ACM Transactions on Database Systems 1*, 1 (1976), 9–36.

[10] Codd, E. F. A relational model of data for large shared data banks. *Communications of ACM 13*, 6 (1970), 377–387.

[11] Copestake, A., and Sparck-Jones, K. Natural language interfaces to databases. *Knowledge Engineering Review 5*, 4 (1990), 225–249.

[12] Fagin, R. A., and Vardi, M. Armstrong databases for functional and inclusion dependencies. *Information Processing Letters 16*, 1 (1983), 13–19.

[13] Kim, J.-Y., Lander, Y. A., and Partee, B. H., Eds. *Possessives and Beyond: Semantics and Syntax*. GLSA Publications and Booksurge LLC, Amherst, MA, USA, Mar. 2005.

[14] Maier, D. *The Theory of Relational Databases*. Computer Science Press, Rockville, USA, 1983.

[15] Martin, P., Appelt, D., and Pereira, F. *Transportability and generality in a natural-language interface system*. Morgan Kaufmann Publishers Inc., Los Altos, CA, USA, 1986, pp. 585–593.

[16] Meng, X., and Wang, S. Nchiql: The chinese natural language interface to databases. In *DEXA'01: Proceedings of the 12th International Conference on Database and Expert Systems Applications* (London, UK, 2001), vol. 2113 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 145–154.

[17] Meng, X., Wang, S., and Wong, K.-F. Overview of a chinese natural language interface to databases: Nchiql. *International Journal of Computer Processing of Oriental Languages 14*, 3 (Sept. 2001), 213–232.

[18] Rappaport, G. C. *The Syntax of Possessors in the Nominal Phrase: Drawing the Lines and Deriving the Forms*. In Kim et al. [13], Mar. 2005, pp. 243–262.

[19] Sagiv, Y. A characterization of globally consistent databases and their correct access paths. *ACM Transactions on Database Systems 8*, 2 (1983), 266–286.

[20] Sheth, A. P., and Larson, J. A. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys 22*, 3 (1990), 183–236.

[21] Storto, G. *Possessives in Context – Issues in the Semantics of Possessive Constructions*. PhD thesis, University of California, Los Angeles, USA, Linguistics, 2003.

[22] Storto, G. *Possessives in Context*. In Kim et al. [13], Mar. 2005, pp. 59–86.

6 Publikációk

Book chapters

[B1] Kardkovács, Z. T. *Szövegbányászat*. TypoTeX, Budapest, Hungary, June 2007, ch. Válaszkereső rendszerek, pp. 217–237.

[B2] Tikk, D., Kardkovács, Z. T., Magyar, G., Babarczy, A., and Szakadát, I. *Intelligent Systems at the Service of Mankind*, vol. 2. UBooks, Augsburg, Germany, Jan. 2006, ch. Natural Language Question Processing for Hungarian Deep Web Searcher.

Articles

[J1] Gajdos, S., Kardkovács, Z. T., and Surányi, G. Deduktív objektum-orientált adatbáziskezelők tervezése és megvalósítása. *Híradástechnika L*, 11 (Nov. 1999), 18–24. Journal on C5.

[J2] Kardkovács, Z. T., Surányi, G., and Gajdos, S. On the integration of large data banks by a powerful cataloguing method. *Periodica Polytechnica 48*, 1–2 (2004), 61–70.

[J3] Kardkovács, Z. T., and Tikk, D. On the transformation of sentences with genitive relations to sql queries. *Data & Knowledge Engineering 61*, 3 (2007), 406–416.

[J4] Kardkovács, Z. T., Tikk, D., and Bánsághi, Z. The ferrety algorithm for the kdd cup 2005 problem. *SIGKDD Explorations 7*, 2 (2005), 111–116.

[J5] Kardkovács, Z. T., Tikk, D., and Bánsághi, Z. A 2005-ös kdd kupa feladatának megoldása a fürkész algoritmussal. *Híradástechnika LXI*, 8 (2006), 50–58. Journal on C5.

[J6] Tikk, D., Kardkovács, Z. T., and Magyar, G. Deep web searcher for hungarian. *Internation Journal on Information Technology 1*, 1–4 (Dec. 2004), 191–197.

[J7] Tikk, D., Kardkovács, Z. T., and Magyar, G. A szavak hálójában: szabadszavas mélyháló-kereső program. *Híradástechnika LX*, 5 (May 2005), 2–8. Journal on C5.

[J8] Tikk, D., Kardkovács, Z. T., and Szidarovszky, F. Szótári névelemek felismerése és morfológiai annotálása. *Híradástechnika LXI*, 1 (2006), 29–34. Journal on C5.

[J9] Tikk, D., Kardkovács, Z. T., and Szidarovszky, F. Voting with a parameterized veto strategy: solving the kdd cup 2006 problem by means of a classifier committee. *SIGKDD Explorations 8*, 2 (2006), 53–62.

Reviewed conference papers

[C1] Surányi, G., Kardkovács, Z. T., and Gajdos, S. Catalogues from a new perspective: A data structure for physical organisation. In 8^{th} *East European Conf. on Advances in Databases and Information Systems, ADBIS 2004* (Budapest, Hungary, 2004), G. Gottlob, A. Benczúr, and J. Demetrovics, Eds., vol. 3255 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 204–214.

[C2] Kardkovács, Z. T. On the transformation of sentences with genitive phrases to sql statements. In *Proceedings of the* 10^{t} *International Conference on Applications of Natural Language to Information Systems (NLDB)* (Alicante, Spain, June 2005), vol. 3513 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 10–20.

[C3] Kardkovács, Z. T., Lejtovicz, E. K., and Kovács, G. Context identification: A relational database approach. In *Proceedings of the* 3^{nd} *Language & Technology Conference* (Poznan, Poland, Oct. 2007), Z. Vetulani, Ed., pp. 211–215.

[C4] Kardkovács, Z. T., Surányi, G., and Gajdos, S. Application of catalogues to integrate heterogeneous data banks. In *Proceedings of On The Move to Meaningful Internet Systems 2003* (Nov. 2003), vol. 2889 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 1045–1056.

[C5] Kardkovács, Z. T., Surányi, G., and Gajdos, S. Ubiquitous access to deep content via web services. In *Web Engineering, International Conference, ICWE 2003* (Oviedo, Spain, 2003), J. M. C. Lovelle, B. M. G. Rodríguez, L. J. Aguilar, J. E. L. Gayo, and M. del Puerto Paule Ruíz, Eds., vol. 2722 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 208–211.

[C6] Kardkovács, Z. T., Surányi, G., and Gajdos, S. Towards building knowledge centres on the world wide web. In 3^{rd} *Int. Conf. on Advances in Information Systems, ADVIS 2004* (Izmir, Turkey, 2004), T. Yakhno, Ed., vol. 3261 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 139–149.

[C7] Kardkovács, Z. T., and Tikk, D. Szintaktikailag elemzett birtokos kifejezések algoritmizált fordítása adott formális nyelvre. In *Magyar Számítógépes Nyelvészeti Konferencia* (2005), D. Csendes and Z. Alexin, Eds., Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 267–276.

[C8] Kardkovács, Z. T., Tikk, D., and Magyar, G. (V)ISA: A model for transforming genitive phrases into sql statements. In *Proceedings of the* 2^{nd} *Language & Technology Conference* (Poznan, Poland, Apr. 2005), pp. 58–62.

[C9] Tikk, D., Bíró, G., Szidarovszky, F., Kardkovács, Z. T., Héder, M., and Lemák, G. Magyar internetes gazdasági tematikájú tartalmak keresése. In *Magyar Számítógépes Nyelvészeti Konferencia* (2006), Z. Alexin and D. Csendes, Eds., Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 3–14.

[C10] Tikk, D., Szidarovszky, F., Kardkovács, Z. T., and Magyar, G. Entity recognizer in hungarian question processing. In *Proceedings of the* 9^{th} *Congress of the Italian Association for Artificial Intelligence* (Milano, Italy, Sept. 2005), vol. 3673 of *Lecture Notes on Artificial Intelligence*, Springer Verlag, pp. 535–546.

[C11] Tikk, D., Szidarovszky, F., Kardkovács, Z. T., and Magyar, G. Named entity recognition in a hungarian nl based qa system. In *Proceedings of the* 14^{th} *International Conference on Information System Developement* (Karlstad, Sweden, Aug. 2005), Kluwer Academic/Plenum Publishers.

[C12] Tikk, D., Kardkovács, Z. T., Andriska, Z., Magyar, G., Babarczy, A., and Szakadát, I. Natural language question processing for hungarian deep web searcher. In 2^{nd} *IEEE International Conference on Computational Cybernetics, ICCC 2004* (Vienna, Austria, Sept. 2004), W. Elmenreich, W. Haidinger, and T. Machado, Eds., pp. 303–309.

[C13] Tikk, D., Kardkovács, Z. T., and Bánsághi, Z. Topic mapping: a tool for finding the meaning of internet search queries. In *INES'06 Proceedings of IEEE Int. Conf. on Intelligent Engineering Systems, 2006* (2006), pp. 227–232.

[C14] Tikk, D., Kardkovács, Z. T., and Magyar, G. The hungarian deep web searcher project. In *International Conf. on Information Technology, ICIT 2004* (Istanbul, Turkey, Dec. 2004), A. Okatan, Ed., pp. 76–79.

[C15] Tikk, D., Kardkovács, Z. T., and Magyar, G. Qa system for hungarian based on deep web search. In *Proceedings of the* 2^{nd} *Romanian-Hungarian Joint Symposium on Applied Computational Intelligence* (Temesvár, Romania, May 2005), pp. 399–410.

[C16] Tikk, D., Kardkovács, Z. T., and Magyar, G. Searching the deep web: The WoW project. In *ISD'2006 Proceedings of the* 15^{th} *International Conference on Advances in Information Systems Development* (Budapest, Hungary, 2006), W. Wojtkowski, W. G. Wojtkowski, J. Zupancic, G. Magyar, and G. Knapp, Eds., Springer Verlag, pp. 493–504.

[C17] Tikk, D., Kardkovács, Z. T., Magyar, G., Babarczy, A., and Szakadát, I. Natural language module of a hungarian deep web searcher. In *IEEE* 4^{th} *International Conference on Intelligent Systems Design and Application* (Budapest, Hungary, Aug. 2004), pp. 73–77.

Other publications

[O1] Lejtovicz, K. E., and Kardkovács, Z. T. Anaforafeloldás természetes nyelvű szövegekben. In *Magyar Számítógépes Nyelvészeti Konferencia* (2006), Z. Alexin and D. Csendes, Eds., Szegedi Tudományegyetem Informatikai Tanszékcsoport, pp. 362–363.

[O2] Kardkovács, Z. T. Deduktív objektumorientált adatbázis-kezelő modellezése és fizikai rétegének hatékony megvalósítása. Master thesis, Budapest University of Technology and Economics, Department of Telecommunications and Telematics, June 2001.

[O3] Kardkovács, Z. T. Az információábrázolás matematikai módszerei. Tech. Rep. MEK-00854, BME Informatikai Központ – Axelero Internet Rt., Magyar Elektronikus Könyvtár, June 2003.

[O4] Kardkovács, Z. T. Fejlett, strukturált információs rendszerek. Tech. rep., BME Informatikai Központ – Axelero Internet Rt., Magyar Elektronikus Könyvtár, May 2003.

[O5] Kardkovács, Z. T., Surányi, G., and Gajdos, S. On the integration of large data banks by a powerful cataloguing method. In *John von Neumann PhD Conference* (Budapest, Hungary, Oct. 2003), pp. 47–50.

[O6] Kardkovács, Z. T., Surányi, G., and Gajdos, S. An axiomatic model for deductive object-oriented databases. In *Proceedings of the* 5^{th} *International Symposium of Hungarian Researchers* (Budapest, Hungary, Nov. 2004), pp. 325–336.

[O7] Tikk, D., Bíró, G., Szidarovszky, F., Kardkovács, Z. T., Héder, M., and Lemák, G. Categorization-based topic-oriented internet search engine. In *HUCI'2006 Proceedings of the* 7^{th} *Symposium of Hungarian Researchers on Computational Intelligence* (Budapest, Hungary, 2006), pp. 233–246.

[O8] Tikk, D., Kardkovács, Z. T., and Magyar, G. Wow: the hungarian deep web searcher. In *Proceedings of the* 5^{th} *International Symposium of Hungarian Researchers* (Budapest, Hungary, Nov. 2004), pp. 135–146.