# Modeling Analogies for Human-Centered Information Systems

Christoph Lofi, Christian Nieke
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{lofi, nieke}@nii.ac.jp

**Abstract:** This paper introduces a conceptual model for representing queries, statements, and knowledge in an analogy-enabled information system. Analogies are considered to be one of the core concepts of human cognition and communication, and are very efficient at conveying complex information in a natural fashion. Integrating analogies into modern information systems paves the way for future truly human-centered paradigms for interacting with data and information, and opens up a number of interesting scientific challenges, especially due to the ambiguous and often consensual nature of analogy statements. Our proposed conceptual analogy model therefore provides a unified model for representing analogies of varying complexity and type, while an additional layer of interpretation models adapts and adjusts the operational semantics for different data sources and approaches, avoiding the shortcomings of any single approach. Here, especially the Social Web promises to be a premier source of analogical knowledge due to its rich variety and subjective content, and therefore we outline first steps for harnessing this valuable information for future human-centered information systems.

## 1 Introduction

Despite the huge success of mobile and Web-based information systems, interaction with such systems still follows *system-centric* interaction paradigms such as hierarchical categorization, list browsing, or keyword searches. *Human-centered* approaches like natural language queries or question answering, which try to emulate the natural interaction of humans with each other, are still few and in the early stages of their infancy. One of the problems which hamper the development of such approaches is that human communication is often ambiguous and carries a lot of implicit information provided by context, common knowledge and interpretation or inference. In this paper, we further this cause by modeling *analogies*, one of the core principles of natural communication, and provide formal foundations for integrating this powerful concept in information systems. This challenging inter-disciplinary research brings together aspects from information systems, social systems, linguistics, and psychology.

Human cognition is largely based on processing similarities of conceptual representations. During nearly all cognitive everyday tasks like e.g., visual perception, problem solving or learning, we continuously perform *analogical inference* in order to deal with new information [1] in a flexible and cross-domain fashion. It's most striking feature is that it is performed on high-level perceptional structures and properties. Moreover, in contrast to formal reasoning or deduction, the use of analogies and analogical inference comes easy

and natural to people. As analogical reasoning plays such an important role in many human cognitive abilities, it has been suggested that this ability is the "core of cognition" [2] and the "thing that makes us smart" [3]. Due to analogy's ubiquity and importance, there is long-standing interest in researching the respective foundations in the fields of philosophy, linguistics, and in the cognitive sciences.

In general, an analogy is a cognitive process of transferring some high-level meaning from one particular subject (often called the *analogue* or the *source*) to another subject, usually called the *target.* When using analogies, one emphasizes that the "*essence"* of source and target is similar, i.e. their most discriminating and prototypical behaviors are perceived in a similar way. As a running example, consider the following analogies: "The Skytree is for Tokyo as the Eiffel Tower is for Paris" (or also "The Skytree is for Tokyo as the Statue of Liberty is for New York"). This simple analogy communicates a lot of implicit information, as for example that the Skytree is an iconic landmark of Tokyo, and a great vantage point. Also, this can lead to very intuitive analogical queries like "is there something like the Eiffel Tower in Tokyo". A more metaphorical analogy example is the famous Rutherford analogy "atoms are like the solar system", an analogy which explained the complex and newly discovered mechanics of the microcosm by pointing out similarities to the well-known celestial mechanics.

In human-centered information systems, there are several interesting and challenging application scenarios for taking advantage of analogies. The prime use case are natural language query interfaces using question answering or verbose queries (e.g. IBM Watson [4] or Apple Siri[1]), but also approaches analyzing user-generated text in Social Media as for example opinion mining, sentiment analysis, or Social Media analytics in general. Also, they can be used to explain suggestions of e-commerce or recommender systems in an easier to understand fashion. For example, in a travel booking portal, the system could use statements like "The Okinawa Islands are the Hawai'i of Japan" to explain the concept to foreign customers agnostic to Japanese holiday locations.

Capturing an abstract notion of similarity is essential for the semantics of analogies. Usually this means that source and target, while being potentially different in many respects (i.e. Tokyo Skytree does not look like the Eiffel Tower at all), *behave* similarly or show similar properties within a larger *context*. Therefore, evaluating and processing analogies relies on the concept of *relational* (or behavioral) *similarity*. Mining these similarities can be realized using different data sources like databases or ontologies, but most notably using user statements in the Social Web and in Social Media.

Our paper aims at paving the way for future research by contributing the following:

- Showcasing analogy-enabled information systems, and discussing the potential data sources with a special focus on the Social Web
- Introducing the *knowledge primitives* required for conceptually modeling analogies for information systems
- Providing a conceptual definition for general *analogy statements,* and a slightly restricted definition more suited for later implementation
- Discussing the semantics of analogy statements, and the generic *operations* required for evaluating them
- Introducing a *generic architecture* for analogy-enabled information systems

[1] http://www.apple.com/ios/siri/

## 2 Towards Analogy-enabled Information Systems

Using analogies in natural speech allows communicating dense information easily and naturally just using few words, as most of the intended semantics will be inferred at the receiver's site. The core semantics are that the analogy source and target *behave* similarly, are *structurally* similar in their context [5], are *perceived* similar [6], or have a *shared high level abstraction* [7]. Understanding analogies is therefore a task requiring powerful cognitive abilities, and computer-based analogy processing is faced with many challenges. From a formal perspective, there have been works interpreting analogies as a special case of *induction* [7], or *hidden inductions* [8] in predicate logics. However, such strict formal modeling neglects the defining characteristic of analogies: their high-level and perceptual nature as most analogies contain a large degree of vagueness and human judgment. Also, considering our previous example query of looking for something equivalent to the Eiffel Tower in Tokyo, two candidates come to mind: the Skytree and Tokyo Tower. Tokyo Tower being similar to the Eiffel Tower can be deduced quite easily, as it is a popular landmark which is also very similar to the Eiffel Tower architecture- and construction-wise. The Skytree on the other hand has only few similarities with the Eiffel Tower, it is newly built and of a completely different design. However, it might be a better answer to the query (i.e. its analogy to the Eiffel Tower is stronger) because its *defining* relationship, i.e. being an iconic and touristically significant landmark, is stronger (and is therefore more similar to the Eiffel Tower conceptually). However, these notions of "more similar", "defining", or "stronger" depend on the common consensus of people and are therefore hard to elicit, or may even change over time or between different groups of persons. Therefore, we argue that any formal model for analogies needs to be able to incorporate these vague concepts in a suitable fashion.

We face these challenges with a two-tier approach, consisting of a *conceptual model*, and one or more respective *interpretation models*. The conceptual model, presented in this paper, allows to represent analogical knowledge, facts, or queries on an abstract level. Working towards an implementation of an analogy-enabled information system, the conceptual model needs additional *interpretation models* defining the operation semantics of the primitives and operations used in the conceptual model. Each interpretation model can therefore have slightly different semantics for operations and concepts like "similarity" or "prototypical relations", which are suited for different scenarios (e.g. defining similarity based on structural aspects in ontologies [9], or relying on the distributional hypothesis [10] for natural language texts).

Current algorithmic approaches only focus on a limited and specialized subset of the analogy semantics, and can be incorporated into our presented two-tier model as specific interpretation models. Most promising seem to be approaches using natural language processing (NLP), and they have been proven to be successful in certain areas of analogy processing. These systems rely on interpreting large (Web-based) text collections [11, 12] and are often tailored to be used with the US-based SAT challenge dataset (part of the standardized aptitude test for college admission) [13, 14]. They are particularly well-suited to capture the consensual nature of perceived similarity by relying on statistics on text written by a large number of (Social) Web users. Besides NLP approaches, there have been experiments with ontologies-based approaches [9], structure-mapping approaches [15], or approaches based on neural networks [16]. Analogical reasoning has also been leveraged for adapting and expanding database schemas [17].

The complexity of interpretation models increases with the desired expressiveness of the analogies the system is supposed to handle, which in turn relies on the semantic distance between the contexts of the analogy source and target. On the one end of this spectrum, we have simple *simile analogies* (not to be confused with the rhetoric figure of speech) which are fairly straightforward, comparing concepts which are very similar with respect to their *attributes* ("A Pony is like a small Horse") but are therefore very limited in their applicability and the amount of transferred implicit information. On the other end we have *metaphors*, which are a very powerful form of analogy that is able to cover large semantic distances between concepts, using very abstract, high-level *structural similarity* [5] ("atoms are like the solar system").

One very common form of analogy is the so-called *4-term analogy model* that consists of two pairs of terms that behave analogous. Our example "The Skytree is for Tokyo as the Eiffel Tower is for Paris" is one of those *4-term analogies*, featuring the pair [Skytree, Tokyo] that is like [Eiffel Tower, Paris]. While the *4-term analogy* is widely used in scientific research on the topic [5, 11, 12], it is actually fairly uncommon in natural speech. Rather than stating a 4-term analogy, people more frequently give statements like: "The Skytree is like the Eiffel Tower" or "Is there something like the Eiffel Tower in Tokyo?". While these statements are based on the same analogy, only parts of the analogy are expressed explicitly. This we refer to as hypocatastasis, meaning that the receiver of the message is supposed to figure out the missing parts of the statement himself, using the general context of the statement, common knowledge and his inference abilities. However, some more complex metaphorical analogies (as e.g. $[atom]::[solar\ system]$) cannot easily be represented in a 4-term form, and are significantly more difficult to process. Therefore, in section 4.3, we will focus more closely on the semantics of 4-term analogies as a trade-off between expressivity and complexity.

## 3 Adapting Semantics and Data Sources

In this section, we will briefly discuss the advantages and shortcoming of different data sources which could be harnessed for our model as future research challenges. For each data source, respective interpretation models are required for capturing the required information on relationships, attributes, and the perceived similarity. Also, keep in mind that analogy-enabled information systems are not limited to a single interpretation model, but can have multiple, specialized interpretation models for different types of analogies, which could even work in parallel with a subsequent combination / voting phase (as for example as in [14, 18]) to avoid the shortcomings of each individual approach.

### 3.1 Relational Databases

Databases (i.e. tabular data) provide precise, explicit information on attribute values of entities (the rows) and also information like their shared class (given by the table itself), e.g. a collection of famous monuments or car models with their specifications. However, they usually lack information on relationships to (abstract) concepts, apart from simple foreign key relationships to other tables. Therefore, interpretation models based on relational databases are suitable for analogies with simpler semantics, as for example similes

and simpler analogies with source and target in closely related contexts. Realizing such a simple interpretation model still poses some interesting challenges:

While most similes are based on similar attributes, the partners are usually not similar in their actual values (like in regular similarity queries), but their values are similar in relation to the respective prototype of their class. For example, the Volkswagen Golf GTI is a very expensive and extremely powerful car compared to other compact cars, just as the Porsche Cayenne GTS is for SUV cars. This leads to the central challenges of discovering the correct classes for comparison (e.g. compact cars vs. SUVs), and the respective prototypical attribute values for obtaining relative statements such as "very expensive" or "extremely powerful". While this can simply be done in the given example by calculating the average for each attribute, the task can become increasingly difficult in other cases. For example, because of its popularity, most people would consider the Apple iPad as the prototypical point of reference for tablet computers. However, its specifications do not match the average product at all. Discovering which values are prototypical might requires additional steps like opinion mining in product reviews [19].

### 3.2 Linked Open Data and Ontologies

The main advantage of Linked Open Data (LOD) sources (e.g. DBpedia, Yago) or various available ontologies is that information on concepts and relationships is explicitly available. The sources are usually cleaned to improve data quality, which, however, may lead to many LOD sources containing only very few (but correct) relationships. For example the entry of "Tokyo Skytree" in DBpedia (one of largest available LOD sources) contains only few attributes like size, location, build date, and very few additional relationships linking for example to the owner company, the architect, or the city district. While the usefulness of current LOD sources for analogy processing is therefore still limited, this will likely improve when more information is incorporated. Furthermore, identifying relevant or prototypical concepts or relationships is more difficult with no quantitative information available, and most approaches which aim at discovering typicality on linked data fall back to text mining (e.g. [20]).

### 3.3 Unstructured Text and the Social Web

Most current approaches for automatically processing analogies are based on processing natural language text, usually crawled from the Web. In contrast to databases and LOD, the Web contains an astonishing amount of information, even on more obscure concepts and entities. However, extracting this information is usually an error-prone task with many challenges. Therefore, current approaches aim at dealing with simplified analogy problems as for example solving multiple-choice analogies from the SAT analogy test data [11–14, 21], basically relying on the distributional hypothesis [10] for heuristically estimating relational similarity. For example in [13], the natural language text snippets relevant to an analogy are obtained via web-search, and then concepts and relationships are identified using predefined extraction patterns. Such approaches are strictly heuristics in their nature, and apply statistics on words with no or only limited further considerations on their actual semantics or type. However, they have shown to deliver good performance for multiple-choice analogy queries, mainly due to the fact that statistically analyzing a large number of web sources allows to grasp perceived similarity and consensus of a large number of people quite well.

Besides purely textual approaches, the Web is also the premier source for establishing prototypicality and similarity measures in structured knowledge-bases, as e.g. described in [20] which discovers typical attributes and relationships for classes. Extracting this type of information from the Social Web (e.g. Blogs, Microblogs like Twitter, product reviews, or even more exotic sources like user-tagged image galleries, or recommender system feedback [22]) promises to be a valuable source for perceptual information.

## 4 Conceptual Model for Analogies

In this section we will first present a high-level design for an analogy-enabled information system. This design is limited by certain practical considerations, and will mostly be suitable for processing analogies which can be represented as 4-term analogies. We will then continue to develop a conceptual model for analogies and related concepts that allows representing the required information and operations.

### 4.1 System Design

One of the challenges of designing an analogy-enabled information system is the ability to combine a range of different data sources and diverse semantics, since we believe that no single approach will be applicable and well suited for every scenario. We therefore propose a system architecture with a layer of interpretation models to be used alternatively or even in parallel, to complement each other. Data sources can include typical knowledge bases or Linked Open Data sources such as DBpedi[2] or Yago[3], containing explicit structured information on entities, instances, their relationships, and properties. Also tabular databases containing structured data on entities and their attributes are possible (e.g. Freebase[4]), but also fully unstructured natural language sources such as Web data and document collections. Therefore, our modeling strives for containing only concepts and operators which can be applied to both structured and unstructured data sources. Basically we follow the semantics of simple structured knowledge bases, but we will sometimes require additional information not necessarily available in an explicit form, and which has to be delivered by the interpretation model (such as how typical or similar certain relations are, a notion which is not stored in most ontologies or knowledge bases). A case study for such an interpretation model is given in section 5.

A visual overview of our general system design is presented in Figure 1. Basically, data sources relying on different data models are transformed into a common shared abstraction, the *knowledge base*. This can be done in a push or pull fashion (i.e., purposefully extracting information for each query by e.g. using on-demand web search, or preemptively extracting a large knowledge base). The primitives for modeling the knowledge base are described in section 4.2. All meta-data, intermediate results as well as final results from analogy processing can be stored in the analogy repository for later re-use. This covers information on similarity or prototypicality, but also discovered analogies and their supporting facts.

[2] http://dbpedia.org/
[3] http://www.mpi-inf.mpg.de/yago-naga/yago/
[4] http://www.freebase.com/

The general semantics of analogy processing are given by our conceptual model, while the specific operative semantics for different data sources, domains, or special cases are provided by respective interpretation models. Please note that in this work, we refrain from discussions about how to parse natural language queries or statements, but assume that the required language processing is provided by a suitable query parser.

### 4.2 Knowledge Base Primitives

In this section, we introduce the basic building blocks for modeling the semantics of analogies in information systems. Please note, our model and architecture is only intended to be used for analogical reasoning. Like in the human thinking process, this allows ignoring several constraints of logical reasoning or formal ontologies, like certain consistency requirements, and will not always give clear, precise answers, as analogies are often ambiguous by nature.

***Concepts $\mathcal{C}$:*** Concepts are our primary modeling primitives, and are identified by their unique label. The set of all concepts is denoted by $\mathcal{C} \subseteq \mathcal{L}$, wheras $\mathcal{L}$ is the set of all unique labels. Several different types of concepts can be encountered and depending on their type, these might require special considerations while designing an interpretation model, resulting in tailored evaluation algorithms or data sources (see section 3). In the following, we distinguish between entities and abstract objects and their special cases classes and prototypes. Heuristics employed in the interpreting model may take advantage of this disambiguation.

- *Entities*: We use entities to represent potential real-world objects which can also be identified by their attribute values, similar to entities in the relational model for databases. This can encompass existing objects as for example the city "Tokyo" or the "Eiffel Tower". Furthermore, entities might also represent more abstract notions which are often used in speech to represent groups of real world instances with the same attributes, as for example the iPhone5 (all actual iPhones have closely similar attribute values, and when used in analogies, no further distinguishing between individual physical objects is usually required). Information on entities is often readily available in Deep Web or LOD data sources, especially in form of structured data or stored in relational databases (for example FreeBase, DBpedia, IMDB, etc.)
- *Abstract Objects*: Abstract objects do not exist in a particular time or place, but rather represent ideas or high-level abstractions as for example "truth", "curiosity", but also

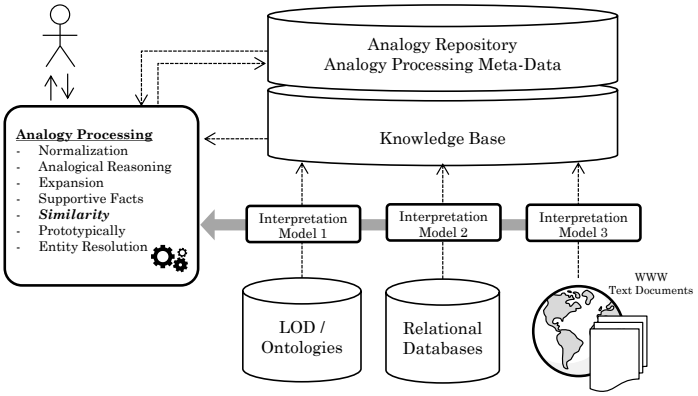


**Figure 1:** General System Overview

abstractions which frequently appear as attribute dimensions like "size" or "costs". Information on abstract objects is usually harder to obtain than information on entities, and they have often only few explicit attribute values.

- *Classes* are a special case of abstract objects, and represent collections of concepts, as for example city, monument, or planet, but also more descriptive classes like "famous monuments loved by tourists which are also towers". One reason for special treatment of classes is due to their implications (especially due to class-subclass relationships) for interpreting algorithms. Class-related information is often explicitly available in web-based ontologies and taxonomies (e.g. ProBase [23]).
- *Prototypes*: Another special case of abstract objects relevant for analogies are prototypes. Prototypes are associated to certain classes, and represent a consensual abstraction of the entities grouped by the classes. For example, the prototypical tablet computer is white and made by Apple. Note that a prototype is not just the average entity, but is usually the prominent representative in peoples' mind. Also, having multiple prototypes per class is possible when perception significantly differs between groups of people. Prototypes play a crucial role when evaluating simile analogies (see section 3.1).

***Attributes $\mathcal{A}$***: Attributes describe the properties of an entity, and can be considered as a labeled relationship between an entity and some alpha-numeric literal values. Attributes only play a minor role for modeling analogies, and we do not further distinguish between different data types for attributes. The set of attributes is given by $\mathcal{A} \subseteq \mathcal{C}\times\mathcal{L}\times\mathcal{I}$, whereas $\mathcal{I}$ is the set of literal values.

***Relationships $\mathcal{R}$***: The set of relationships further describes the relations and interactions between concepts. Each relationship is labeled, and the label represents certain real-world semantics. As the knowledge base does not have to follow a strict vocabulary or schema, one challenge with respect to relationships is that there may be multiple relationships with different labels describing similar real-world semantics.

The set of relationships $\mathcal{R}$ is given by $\mathcal{R} \subseteq \mathcal{C}\times\mathcal{L}\times\mathcal{C}$. Furthermore, we define a function $r\colon \mathcal{C}\times\mathcal{C} \rightarrow (\mathcal{C}\times\mathcal{L}\times\mathcal{C})$ returning all relationships between two given concepts.

### 4.3 Analogons, General Analogies and 4-Term Analogies

An analogy is a high-level comparison between *source* and *target*. Both are represented as *analogons* and the general form of all analogies can be denoted as $A :: B$ (with $A$ being the source and $B$ being the target analogon). In general, analogons represent a set of concepts (usually with one dominant core concept while the others are related) and, implicitly, the relations between them. We formally define the set of all analogons as a subset of the power set of all concepts: $\mathcal{A}g_{Full} \subseteq \mathcal{P}(\mathcal{C})$. This definition is close to one of the dominant views in cognitive sciences, where analogies are often described as high-level structural mappings between two complexes mental representations [5] (concepts and relationships).

However, this complex cognitive model is difficult to realize, and therefore we will focus on the less complex special case of *4-term analogies* as the basic and *canonical form* for representing analogy statements and queries in our model. Of course, an adaption to the more general case is possible, but would strongly increase the required effort for the problem's presentation and is usually beyond the reach of most of the current and still prototypical implementations. In the case of 4-term analogies, only a restricted set of those analogons which contain exactly two concepts is used: $\mathcal{A}g_4 \subseteq \mathcal{C}\times\mathcal{C} \subseteq \mathcal{A}g_{Full}$.

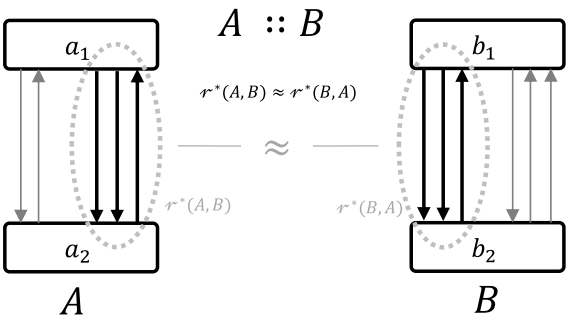


**Figure 2:** Intended semantics of a 4-term analogy statement: An analogy holds true when the relevant relationships between the concepts of the analogons are sufficiently similar

Using analogons, *4-term analogy* statements (in the following also referred to as *canonical* analogy statements) can be expressed as:

$$A :: B \text{ with } A, B \in \mathcal{Ag}_4 \quad \text{or} \quad [a_1, a_2] :: [b_1, b_2] \text{ with } a_1, a_2, b_1, b_2 \in \mathcal{C}$$

We also allow incomplete analogons with ? representing the missing concept. The shorthand notation $[a_1]$ represents $[a_1, ?]$. We refer to the set of all canonical analogy statements as $\mathcal{Analogy}_4$, which is a subset of general analogies $\mathcal{Analogy}_{Full}$.

### 4.4 Semantics of Analogy Statements

In order to define the semantics of 4-term analogy statements, several additional operators are required on the conceptual level. Analogously to logical languages and their respective interpretations, our *interpretation models* will provide the operative semantics of these operators and the involved modeling primitives. Therefore, they also decide which analogy statements actually hold true and which don't. Unfortunately, unlike logic-based systems or relational databases, these operators can only be defined heuristically and have to mimic the imprecise and vague semantics of analogies, trying to grasp the consensual opinions on concepts and their relations.

Basically, the intended semantics of a canonical analogy statement is that the statement holds true if all *relevant relationships* (in the context of the current analogy) between the concepts of each analogon are *closely similar*. Furthermore, there can be stronger analogies (i.e. good analogies which people usually consensual agree to, i.e. the Skytree-Eiffel Tower analogy) or weaker analogies (i.e. analogies people often do not agree with, like $[university, life] :: [strawberry, cake]$). This is reflected by applying an *interpretation function* $\mathcal{I}$, given by an interpretation model, to an analogy statement. The interpretation function will return a numerical value between 0 (weak) and 1 (strong) representing the respective strength of the analogy (analogously to interpreting logical statements which evaluate to true or false), i.e. $\mathcal{I}$: $\mathcal{Analogy}_4 \rightarrow [0,1]$. An analogy *holds true* when this value is above a threshold specific to each interpretation model. Interpretation models are further discussed in section 3.

$r^*$ **(Relevant Relationships):** The first conceptual operation required to formally define the analogy is retrieving the set of relevant relationships, which is given by:

$$r^*(A, B) = r \text{ with } A, B \in \mathcal{Ag}_4 \text{ and } r \subseteq \mathcal{R}$$

This function returns those relationships between $a_1$ and $a_2$ (with $[a_1, a_2] = A$) which are relevant to the analogy $A :: B$. As we consider analogies being symmetric, i.e. $\mathcal{I}(A :: B) = \mathcal{I}(B :: A)$, $\mathcal{r}^*(B, A)$ can be used to obtain all those relationships between $b_1$ and $b_2$ relevant for the analogy $A :: B$.

Unfortunately, to realize these intended semantics in an interpretation model one has to rely on heuristic assumptions for "relevant for the analogy" (please note that humans interpreting analogies also heuristically infer these relationships individually for each analogy). The simplest heuristic is to assume that all relationships between the concepts of an analogon are relevant, which will result in weaker semantics but might be sufficient in certain restricted scenarios. A practically feasible trade-off between complexity and expressivity is relying on (proto-) typicality, i.e. assuming that those relationships are relevant for the analogy which are typical for the concept pair of the analogon (i.e. for [Skytree, Tokyo], being an iconic landmark). For this task, approaches for capturing property or relationship typicality as for example [19, 20] can be adapted.

$\sigma$ **(Relationship Set Similarity):** Having defined the sets of relevant relationships, the next step is to define a notion of similarity between these sets. We denote the *relationship similarity function*:

$$\sigma: R_1 \times R_2 \rightarrow [0,1] \text{ with } R_1, R_2 \subseteq \mathcal{R}$$

The function describes the similarity between two sets of relationships (with 1.0 representing maximal similarity) and can be used to describe the strength of an analogy when applied to the sets of relevant relationships:

$$\mathcal{I}(A :: B) = \sigma(\mathcal{r}^*(A, B), \mathcal{r}^*(B, A))$$

We further denote: $R_1 \approx R_2$ as the relationship sets $R_1, R_2$ being *sufficiently similar*, i.e. their similarity is above an interpretation model specific threshold:

$$R_1 \approx R_2 \equiv R_1, R_2 \subseteq \mathcal{R} \wedge \sigma(R_1, R_2) > threshold$$

Again, the operative semantics of this notion and the threshold have to be established by an interpretation model. Here, most heuristics for set similarity will have to rely on the similarity between two single relationships, which is discussed in the next section.

$\boldsymbol{\sigma_R}$ **(Relational Similarity):** Similarity in the context of analogies does not refer to attribute similarity, but to the more challenging concept of *relational similarity.* We will denote relational similarity between two relationships as a function returning a value ranging from 0 to 1, with 1 representing maximal similarity:

$$\sigma_R: R \times R \rightarrow [0,1]$$

One heuristic for approaching this problem is the distributional hypothesis [10] from linguistics, which claims that words frequently occurring in the same context also have similar meanings. An implementation of this heuristic can be found in [13]. Another heuristic approach relying on pattern extraction is given in [24]. Also, approaches based on crowd-sourcing can provide valuable input [18]. Still, developing effective heuristics for relational similarity needed for interpretation models remains as one of the core challenges of future research - poor implementation of similarity might capture only simple analogies, while a superior heuristic interpretation model may cover even more complex metaphors.

Please note that especially for similes, also attribute values may play a role. This is still covered by our model, as the semantics of such analogies are that attribute values are similar relative to another concept or in a certain context (e.g. compared to a prototype or

**Table 1**: Different Representations of Analogies

| | Natural Communication | Formal Example |
|---|---|---|
| **Implicit Analogy** | "The Skytree is analog to / like the Eiffel tower" | $[a_1] :: [\mathrm{b}_1]$ $\equiv [a_1, ?] :: [b_1, ?]$ |
| **Explicit Analogy** | "The Skytree is to Tokyo what the Eiffel tower is to Paris" | $[a_1, a_2] :: [b_1, b_2]$ |
| **Expansion** | Inferred by receiver: List of relevant relationships and their similarity, representing the transferred knowledge | $\{\mathcal{I}(A :: B) = s,$ $\sigma_R(r_{a,1}, r_{b,1}) = s_1, \dots\}$ |
| **Supporting Facts** | Facts supporting the decisions on relevance and similarity, e.g.: "Tokyo is a city", "Paris is a city" | $\{isCity(a_2), isCapital(b_2),$ $isA(City, Capital), \dots\}$ |

reference concept), and are therefore captured by relational similarity. For example, a Porsche 911 is *expensive* for a sport car, as is the Mercedes Benz S-Class for a sedan. However, mining the correct relationships from attribute values is its own challenge (see 3.1). A graphical summary of our intended analogy semantics can be found in figure 3.

### 4.5 Normalizing Analogies, Expansions, and Supportive Facts

As indicated before, analogies are usually given in natural speech and are therefore often not fully explicit. In our model, we propose a *normalization* of analogies by transforming an *implicit* analogy into its *explicit canonical form*, i.e. a complete 4-term analogy. This step requires to find the *implicit* information in a statement like "The Skytree is analog to the Eiffel tower" to normalize it into the explicit form "The Skytree is to Tokyo what the Eiffel tower is to Paris". Since the explicit statement contains more information than the implicit one, this transformation can obviously not be performed by applying a strict set of rules, but is in itself a non-trivial cognitive task relying on heuristic inference.

One simple heuristic for normalization, i.e. inferring the missing concepts, could be realized as follows: Given an implicit analogy statement $[a_1] :: [b_1]$ (equivalent to $[\mathrm{a}_1, ?_1] :: [\mathrm{b}_1, ?_2]$), the values of $?_1$ and $?_2$ can be obtained by inspecting all possible assignments of $?_1, ?_2$ (i.e. those concepts which have a relationship with $a_1$ or $b_1$), and selecting that pair for which the prototypical relationships are similar. This task is closely related to evaluating an explicit analogy statement, and the same heuristics provided by an interpretation model for relevance and similarity can be reused. The careful inclusion of crowd-sourcing could also be beneficial.

The *analogy expansion* and *supportive facts* are optional features of our model. Basically, after evaluating an analogy statement or query, they provide additional information on the evaluation process. The *expansion* represents the information that is implicitly transferred by the analogy, i.e. the relevant relations existing in both analogons ("isIconicLandmark"), with their similarity. The *supportive facts* are facts from the knowledge base used to derive the similarity and relevance of the relations. In summary, an analogy is true because of the similar and relevant relations shown in the expansion, which were derived using the data in the supportive facts. You can see an overview of the different proposed representations of analogies in Table 1.

### 4.6 Analogy Queries

To query an analogy based information system, we propose to state the analogy with all unknown concepts or parts replaced by a '?'. The answer returned by the system would be all analogy statements in canonical form that fulfill the request, optionally including the extension and the supporting facts, or an empty result set if no analogy could be found. This process is very similar to normalizing statements. In the following, we present some examples of possible analogy queries:

- $?\,[a_1, a_2] :: [b_1, b_2]$: This statement checks if the given analogy statement is true (or the result set is empty) and retrieves its extension and supporting facts.
- $?\,[a_1, a_2] :: [?\,, b_2]$: This statement can be used to find the missing concept in a 4-term analogy. ("What is to Tokyo like the Eiffel Tower is to Paris?")
- $?\,[a_1, a_2] :: ?$ : This statement can be used to find possible matches to an analogon. This is basically the non-multiple-choice version of the SAT analogy challenge. The SAT challenges are significantly easier to solve, as they just test 5 candidate analogons and decide for the one with the highest similarity.
- $?\,[a_1, ?\,] :: [b_1, ?\,] \equiv [a_1] :: [b_1]$: This statement can be used to request the explicit form of an implicit analogy.

## 5 Case Study: Mining Analogies from the Social Web

In this section, we briefly show a case study outlining first steps towards a technical implementation of an analogy interpretation model based on mining the Social Web, and therefore demonstrate the real-world applicability of our proposed model. This study is discussed in closer detail in our later works, and is showcased here to highlight the feasibility and applicability of analogy-enabled information systems. We focus on analogies between locations as e.g. used in travel and tourism systems, e.g. "West Shinjuku (a district of Tokyo) is like Lower Manhattan". During the course of the study, we used Web search with Hearst-like patterns to obtain a set of 22.360 Web documents retrieved mostly from various discussion forums and blogs. Using automated filtering heuristics and crowdsouring-based manual filtering we extracted a Gold dataset with short text snippets containing such analogies. Then, we designed, trained, and evaluated supervised learning models with rich feature sets to *recognize analogy statements automatically*, allowing us to substitute manual crowd filtering with automated techniques. Our best performing feature set based on subsequence patterns derived from PrefixSpans [25] resulted in a precision of 0.92 with recall of 0.88, clearly demonstrating that it is possible to extract analogical statements from the Social Web reliably in an automated fashion.

From here, the next steps are to actually extract the information required to implement the interpretation model: "relevant relationships", "prototypical relationships", and "perceived similarity" between relationships. This task is particularly promising when working on Social Web data, as user's often explicitly explain the analogy they used by providing why they believe it holds true. Consider one of the text snippets from our Gold dataset: "Tokyo, like Disneyland, is sterile. It's too clean and really safe, which are admirable traits, but also unrealistic. Tokyo is like a bubble where people can live their lives in a very naive and enchanted way because real problems do not exist." Here, the speaker clearly provides that in her opinion, Tokyo´s most prototypical properties are that it is very

clean and safe, and that it shares these properties with Disneyland. However, in order to use such statements for building an analogy knowledge base as described in the previous chapters, extraction models have to be designed and trained to locate the properties compared in this natural language statement. This task is actually quite closely related to trigger detection in NLP exploiting analysis of the word-dependency graph representation of the sentence. This problem, and of course aggregating the (potentially subjective and even conflicting) information obtained from multiple statements authored by different users remains as a challenge to be solved by current research in progress.

## 6 Summary & Outlook

During the course of this paper we outlined a generic conceptual design for an analogy-enabled information system that can be adapted to different data sources and operative semantics using one or several interpretation models. We continued with an overview on different data sources, and highlighted their challenges and prospects and provided some insights for designing respective interpretation models, including references to existing work. We then gave a formal model for analogies, including an explicit canonical form, and related problems as e.g. normalization of implicit analogies. We further defined the semantic operations needed to perform analogical reasoning and query the system. Also, we provided a brief survey outlining how an interpretation model based on text crawled from the Social Web can be realized. In our future works we plan to complete this interpretation model, and also provide several additional interpretation models based on alternative data sources for selected scenarios for a complete multi-model analogy-enabled information system, able to adapt to situations where a single model would fail.

## References


1. Gentner, D., Markman, A.B.: Structure mapping in analogy and similarity. American Psychologist. 52, 45–56 (1997).
2. Hofstadter, D.R.: Analogy as the Core of Cognition. The Analogical Mind. pp. 499–538 (2001).
3. Gentner, D.: Why We're So Smart. Language in Mind: Advances in the Study of Language and Thought. pp. 195–235. MIT Press (2003).
4. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. AI Magazine. 31, 59–79 (2010).
5. Gentner, D.: Structure-mapping: A theoretical framework for analogy. Cognitive science. 7, 155–170 (1983).
6. Chalmers, D.J., French, R.M., Hofstadter, D.R.: High-level perception, representation, and analogy: A critique of artificial intelligence methodology. Journal of Experimental & Theoretical Artificial Intelligence. 4, 185–211 (1992).
7. Shelley, C.: Multiple Analogies In Science And Philosophy. John Benjamins Pub. (2003).
8. Juthe, A.: Argument by Analogy. Argumentation. 19, 1–27 (2005).
9. Forbus, K.D., Mostek, T., Ferguson, R.: Analogy Ontology for Integrating Analogical Processing and First-principles Reasoning. Nat. Conf. on Artificial Intelligence (AAAI). , Edmonton, Alberta, Canada (2002).

10. Harris, Z.: Distributional Structure. Word. 10, 146–162 (1954).
11. Ishizuka, M.: Exploiting macro and micro relations toward web intelligence. 11th Pacific Rim Int. Conf. on Trends in Artificial Intelligence (PRICAI). , Daegu, Korea (2010).
12. Turney, P.: A Uniform Approach to Analogies, Synonyms, Antonyms, and Associations. 22th Int. Conf. on Computational Linguistics (COLING). , Beijing, China (2008).
13. Bollegala, D.T., Matsuo, Y., Ishizuka, M.: Measuring the similarity between implicit semantic relations from the web. 18th Int. Conf. on World Wide Web (WWW). , Madrid, Spain (2009).
14. Turney, P.D., Littman, M.L., Bigham, J., Shnayder, V.: Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. Int. Conf. on Recent Advances in Natural Language Processing (RANLP) (2003).
15. Gentner, D., Gunn, V.: Structural alignment facilitates the noticing of differences. Memory & Cognition. 29, 565–77 (2001).
16. Hummel, J.E., Holyoak, K.J.: Relational Reasoning in a Neurally Plausible Cognitive Architecture. An Overview of the LISA Project. Current Directions in Psychological Science. 14, 153–157 (2005).
17. Breitman, K.K., Barbosa, S.D.J., Casanova, M.A., Furtado, A.L.: Conceptual modeling by analogy and metaphor. ACM Conf. in Information and Knowledge Management (CIKM). , Lisabon, Portugal (2007).
18. Lofi, C.: Just ask a human? – Controlling Quality in Relational Similarity and Analogy Processing using the Crowd. CDIM Workshop at Database Systems for Business Technology and Web (BTW). , Magdeburg, Germany (2013).
19. Selke, J., Homoceanu, S., Balke, W.-T.: Conceptual Views for Entity-Centric Search: Turning Data into Meaningful Concepts. Computer Science: Research and Development. 27, 65–79 (2012).
20. Lee, A., Wang, Z., Wang, H., Hwang, S.: Attribute Extraction and Scoring: A Probabilistic Approach. Int. Conf. on Data Engineering (ICDE). , Brisbane, Australia (2013).
21. Davidov, D.: Unsupervised Discovery of Generic Relationships Using Pattern Clusters and its Evaluation by Automatically Generated SAT Analogy Questions. Ass. for Computational Linguistics: Human Language Technologies (ACL:HLT). , Columbus, Ohio, USA (2008).
22. Selke, J., Lofi, C., Balke, W.-T.: Pushing the Boundaries of Crowd-Enabled Databases with Query-Driven Schema Expansion. 38th Int. Conf. on Very Large Data Bases (VLDB). pp. 538–549. in PVLDB 5(2), Istanbul, Turkey (2012).
23. Wu, W., Li, H., Wang, H., Zhu, K.: Probase: A Probabilistic Taxonomy for Text Understanding. SIGMOD Int. Conf. on Management of Data. , Scottsdale, USA (2012).
24. Nakashole, N., Weikum, G., Suchanek, F.: PATTY: A Taxonomy of Relational Patterns with Semantic Types. Int. Conf. on Empirical Methods in Natural Language Processing (EMNLP 2012). , Jeju Island, Korea (2012).
25. Pei, J., Han, J., Mortazavi-asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. IEEE Computer Society. (2001).