Knowledge-Based Systems 29 (2012) 20–30

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Evaluation of analogical proportions through Kolmogorov complexity

Meriam Bayoudh [b,1], Henri Prade [a], Gilles Richard [a,*]

[a] IRIT, 118 Route de Narbonne, 31062 Toulouse Cedex 9, United Kingdom
[b] Centre IRD de Guyane, Route de Montabo BP165, 97323 Cayenne CEDEX, France

ARTICLE INFO

Article history:
Available online 18 July 2011

Keywords:
Analogical proportion
Kolmogorov complexity
Common sense analogies
Search engine
Google

ABSTRACT

In this paper, we try to identify analogical proportions, i.e., statements of the form "*a* is to *b* as *c* is to *d*", expressed in linguistic terms. While it is conceivable to use an algebraic model for testing proportions such as "2 is to 4 as 5 is to 10", or even such as "read is to reader as lecture is to lecturer", there is no algebraic framework to support statements such as "engine is to car as heart is to human" or "wine is to France as beer is to England", helping to recognize them as meaningful analogical proportions. The idea is then to rely on text corpora, or even on the Web itself, where one may expect to find the pragmatics and the semantics of the words, in their common use. In that context, in order to attach a numerical value to the "analogical ratio" corresponding to the phrase "*a* is to *b*", we start from the works of Kolmogorov on complexity theory. This is the basis for a universal measure of the information content of a word *a*, or of a word *a* with respect to another one *b*, which, in practice, is estimated in a statistical manner. We investigate the link between a purely logical, recently introduced view of analogical proportions and its counterpart based on Kolmogorov theory. The criteria proposed for testing candidate proportions fit with the expected properties (symmetry, central permutation) of analogical proportions. This leads to a new computational method to define, and ultimately to try to detect, analogical proportions in natural language. Experiments with classifiers based on these ideas are reported, and results are rather encouraging with respect to the recognition of common sense linguistic analogies. The approach is also compared with existing works on similar problems.

1. Introduction

Despite its heuristic status, analogical reasoning is a commonly used form of reasoning which has the ability to shortcut long chains of classical deductions, while often reaching the same conclusions. It is largely accepted that analogy is the basis for creativity as it puts different paradigms into correspondence (see [16,35,14]). Analogical reasoning is based on the human ability to identify "situations" or "problems" *a* and *c*, and then "deduce" that if *b* is a solution for the problem *a*, then some *d*, whose relation to *c* is similar to the relation between *a* and *b*, might be a solution for problem *c*. Such a relation involving 4 items *a*, *b*, *c*, *d* is called an analogical proportion, or analogy for short, usually denoted $a{:}b{::}c{:}d$ and should be read "*a* is to *b* as *c* is to *d*". Algebraic frameworks for giving concise definitions of analogical proportions have been deeply investigated in [37] in the last past years. For instance, when the universe is the set $\mathbb{R}$ of real numbers, the truth of $a{:}b{::}c{:}d$ is interpreted as $a - d = b - c$, justifying "2 is to 4 as 5 is to 10". Another example now involving sequences of bits could be 01 is to 10 as 11 is to 00 just because 01 and 10 does not share any bit and this is also the case with 11 and 00. In [28,30], a complete logical framework has been developed, mainly Boolean-oriented, i.e. where the underlying universe is $\mathbb{B} = \{0, 1\}$ or isomorphic to $\mathbb{B}$. In the field of artificial intelligence, analogy-discovering programs have been designed for specialized areas where there exists an underlying minimal algebraic structure.

Natural language analogies like "engine is to car as heart is to human" or "wine is to France as beer is to England" are more at a linguistic or conceptual level: a simple mathematical structure is missing to cope with such proportions. Sowa's conceptual graphs (CG) offer an appealing framework for representing concepts: core knowledge can be encoded using CG, and then with the help of a structured linguistic database (like e.g. WorldNet), we could discover analogies as with VivoMind analogy engine [35]) for instance.

There is another option coming from the works of Gentner ([13,14] about the so-called structure mapping theory (SMT), implemented in the structure mapping engine (SME) with [11]. This way to proceed allows the author to exhibit high level analogical proportions: for instance, the analogical proportion "planets are to the sun as electrons are to the atom's kernel" is coming from the mapping between a representation of the solar system and a representation of the Bohr model of atom. Obviously, this can only

* Corresponding author.
E-mail addresses: meriam.bayoudh@ird.fr (M. Bayoudh), prade@irit.fr (H. Prade), richard@irit.fr (G. Richard).

[1] On leave from IRIT, presently at Université des Antilles et de la Guyane at Cayenne.

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.06.022

be done with the help of a costly high level hand-coded representation. And this is exactly what we want to avoid here! In the field of Computational Linguistics, the works of Turney et al. [42,41] relying on corpus-based techniques to learn semantic features like analogies, synonyms, antonyms and associations, are very successfully and we will devote Section 5 to investigate this approach and to compare with the one we propose.

But, let us carry on with our ideas now. In [27], a method dealing with natural language analogies but avoiding any pre-coding of the universe has been developed. The main idea is that each word a carries an "information content" that is formally defined via its Kolmogorov complexity, $K(a)$, which is an ideal natural number. In order to build up an effective implementation, this number has to be estimated. Thanks to the works of Solomonoff [34], it appears that $K(a)$ can be related to the probability of a to "appear". Thus, applying a kind of reverse process, we start from a probability distribution to estimate the Kolmogorov complexity. Among the candidates to provide a probability distribution over the set of English words, the World Wide Web is a strong one. Considering Google as a web mining engine, it is an easy game to get for each word (in our case, a concept representation) its frequency and to consider it as a probability to appear in a document. Then we are done with the estimation of the Kolmogorov complexity of a word: applying our definitions involving only the complexity of a, b, c and d, we can now check if $a{:}b{::}c{:}d$ holds or does not hold. It appears that the proposed definitions are rather consistent with a sample of well-agreed analogies, as we shall see.

Obviously, the web is a relatively dynamic corpus and we could imagine to improve our works within a more homogeneous database, where in some sense, noise has been filtered. Starting from our previous works, we first re-implement a classifier using a structured database coming from the US National Institute of Standards and Technology (NIST) TREC Document Databases.[2] Then a careful examination of our results leads to propose other options, bridging the gap between a purely Boolean view and a Kolmogorov-based definition.

Our paper is organized as follows: the next section starts from an informal analysis of the core concepts underlying an analogical proportion leading to the well agreed axioms defining this proportion. We also provide the Boolean interpretation of such a proportion and highlights the properties we expect to be satisfied in another context. In Section 3, we switch to natural language analogies, briefly recalling the main principles of Kolmogorov complexity theory and its companion concept known as the universal distribution. We show how to use it to provide different practical definitions for analogical proportion between concepts represented as words, highlighting the link with the logical setting described in Section 2. In Section 4, we examine the results we get through diverse sets of experimentations and we show that they bridge the apparent gap between the Boolean framework and the complexity-based framework. Finally we survey related works and conclude. Sections 5 and 6 provide a comparative discussion of the proposed model with another approach developed in computational linguistics, at the methodological level and on a preliminary experiment. This paper is a fully revised and substantially expanded version of a conference paper [3].

2. Analogical proportions: a logical view

An analogical proportion[3] can be considered as a relation involving 4 items and satisfying some basic axioms which are supposed to capture its essence and that we recall below. Let us start with an informal analysis of the core concepts underlying this relation.

2.1. Brief analysis

In order to transfer knowledge, analogical reasoning considers two situations in parallel and compare them by putting them into correspondence. In the structure mapping theory terminology [11], the output of this process would be the so-called "mapping function". Here, we want to stick to a simpler context where each situation involves only two entities or items, say a, c on the one hand, and b, d on the other hand. The comparison then bears on the pair a and b, and on the pair c and d. This naturally leads to consider two kinds of properties:

- what is common in terms of properties to a and b: let us denote it $com(a,b)$,
- and what is specific to a and not shared by b: we denote it $spec(a,b)$.

Due to the intended meaning of com and $spec$, it is natural to assume $com(a,b) = com(b,a)$ but in general, we cannot assume $spec(a,b) = spec(b,a)$: $spec(a,b) \neq spec(b,a)$ is more realistic. With this view,

- a is represented by the pair $(com(a,b), spec(a,b))$
- b is represented by the pair $(com(a,b), spec(b,a))$

while

- c is represented by the pair $(com(c,d), spec(c,d))$
- d is represented by the pair $(com(c,d), spec(d,c))$

Then, an analogical proportion between the 4 items, expressing that *a is to b as c is to d* amounts to state that the way a and b differ is the same as the way c and d differ, namely using our notation:

$$spec(a,b) = spec(c,d) \quad \text{and} \quad spec(b,a) = spec(d,c)$$

assuming symmetry in the way the parallel is done. This simple informal observation highlights two expected properties:

- a is to b as a is to b and
- if a is to b as c is to d then c is to d as a is to b (due to the symmetry of the = operator)

Going a little bit deeper in this informal analysis, we can also observe above that since $spec(a,b) = spec(c,d)$, it means that a differs from c through the properties of a shared with b previously denoted $com(a,b)$, and it is the same for b with respect to d. This amounts to write $spec(a,c) = spec(b,d)$ since they are both equals to $com(a,b)$. A symmetric reasoning leads to $spec(c,a) = spec(d,b)$, which together which the previous equality exactly mean a is to c as b is to d. We retrieve here the central permutation postulate that most authors associate with analogical proportion together with the symmetry postulate already mentioned. We have thus retrieved the 3 characteristic properties usually requested for a proper definition of analogical proportions. It is time now for a formalization.

2.2. Formal setting

The best option is to consider a first order setting where a, b, c, d are variables and A denotes a quaternary relation. A is an analogical proportion when it satisfies the following axioms:

- $A(a,b,a,b)$ (identity)
- $A(a,b,c,d) \Rightarrow A(c,d,a,b)$ (symmetry)

[2] http://www.nist.gov/srd/nistsd22.htm.

[3] From time to time in the remaining of the paper, the word 'analogy' will be used as a shortcut for 'analogical proportion'.