Context-dependent relevant descriptions

Jean-Louis Dessalles
Telecom Paris
jl @ dessalles.fr
www.dessalles.fr

TELECOM Paris

INSTITUT POLYTECHNIQUE DE PARIS

REMI: Mining Intuitive Referring Expressions on Knowledge Bases

Luis Galárraga
Inria
luis.galarraga@inria.fr

Julien Delaunay
University of Rennes I
juliendelaunay35000@gmail.com

Jean-Louis Dessalles
Télécom ParisTech
dessalles@telecom-paristech.fr

ABSTRACT

A *referring expression* (RE) is a description that identifies a set of instances unambiguously. Mining REs from data finds applications in natural language generation, algorithmic journalism, and data maintenance. Since there may exist multiple REs for a given set of entities, it is common to focus on the most concise and informative (i.e., intuitive) ones. We present REMI, a method to mine intuitive REs on large knowledge bases. Our experimental evaluation shows that REMI finds REs deemed intuitive by users. Moreover we show that REMI is several orders of magnitude faster than an approach based on inductive logic programming.

1 INTRODUCTION

The approach in [6] overcomes this limitation to some extent, by allowing users to provide a ranking of preference for the attributes used in the description. Nevertheless, providing such a ranking can be tedious for KBs with thousands of predicates.

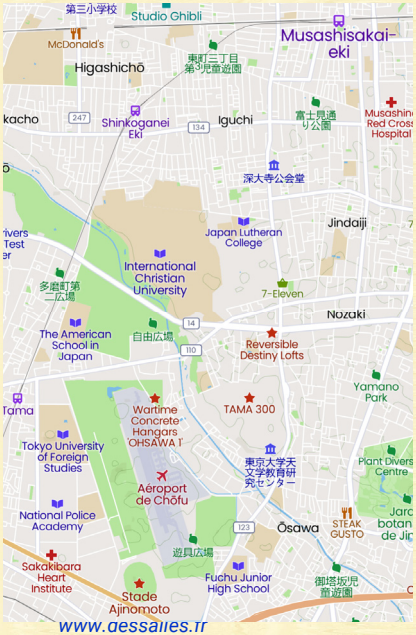
We tackle the aforementioned limitations with a solution to mine intuitive REs on large KBs. How to use such REs is beyond the scope of this work, however we provide hints about potential use cases. In summary, our contributions are:

- A scheme based on information theory to quantify the intuitiveness of entity descriptions extracted from a KB.
- REMI, an algorithm to mine intuitive REs on large KBs. REMI extends the state-of-the-art language bias for REs and allows for expressions such as $mayor(x, y) \wedge party(y, Socialist)$. This design choice increases the chances of finding intuitive REs

Context-dependent relevant descriptions

- **Idea #1**
  - **Retrieve an entity (e.g. from Yago)**
    - *e.g. "Musashisakai"*
  - **Find determinations**
    - *absolute/relative location*
    - *RDF (?) predicates (population, country, history)*
  - **Compute the relevance of determinations**
    - *In number of bits*

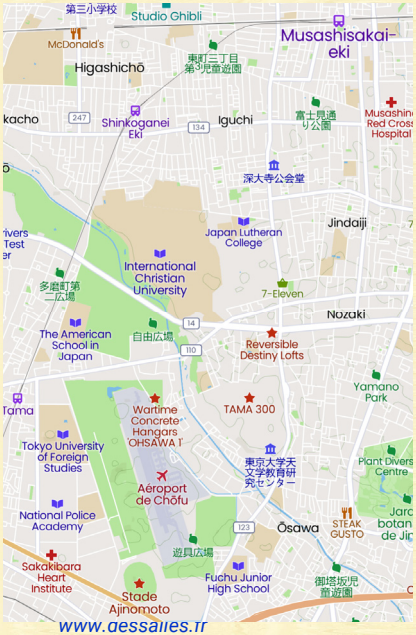
Musashisakai



Musashi-Sakai Station is served by the JR East Chūō Main Line,
and is also the northern terminus of the short Seibu Tamagawa Line.
It is not a major transfer station,
and only local (all-stations) trains on the Chūō Line stop at Musashi-sakai.

The JR station opened on 11 April 1889

Musashisakai



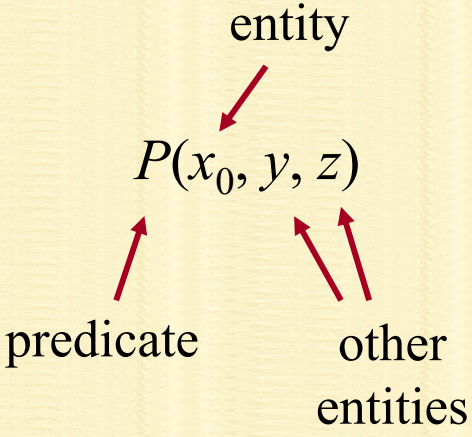
Musashino (武蔵野市, *Musashino-shi*)
is a city
located in the western portion of Tokyo Metropolis, Japan.
As of 1 February 2016, the city had an estimated population of 143,868,
and a population density of 13,090 persons per km².
Its total area is 10.98 square kilometres (4.24 sq mi).[1]
Based on the 2015 Kanto Ranking, Musashino was the 5th most desirable place to live in Central Japan.
Popular attractions in Musashino include Kichijōji;
a residential and shopping neighborhood with malls such as Atre Kichijoji,
recreational areas such as Inokashira Park, Musashino Chuo Park,
Musashino Municipal Athletic Stadium and Musashino Sports Complex.

Relevance

- **Relevant determinations make the entity simple**
  - **Counter-Example**
    - the city had an estimated population of 143,868
  - **Example**
    - Musashino was the 5th most desirable place to live in Central Japan

Relevance

- ***Relevant determinations make the entity <u>simple</u>***



$$C(x_0) \leq C(P) + C(y) + C(z) + \log_2 |\{x;\ P(x, y, z)\}|$$

Relevance

- ***Relevant determinations make the entity <u>simple</u>***

trade-off

$$C(P) + C(y) + C(z)$$

$$+ \log_2 |\{x;\ P(x, y, z)\}|$$

Relevance

- **Relevant determinations make the entity <u>simple</u>**

trade-off

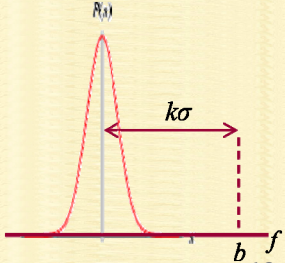
$$C(P) + C(y) + C(z)$$
$$+ \log_2 |\{x;\ P(x, y, z)\}|$$

chain rule $\quad C(x_0) \leq C(y) + C(\mathrm{x}|y)$

Complexity computations

- **Proxies**
  - **$\log_2$(#rank in list)**
    - "5th most desirable place" → 3 bits
    - #hits in search engine
  - **2 × $\log_2$(distance/size)**
    - "20 km west of downtown Tokyo"
    - $C(L) + \log_2(x/L)$
  - **Atypicity**
    - $C(x) = \log_2 N - A(k)$; $A(k) \sim k^2$

$$C(x)$$



Context-dependent relevant descriptions

- **Idea #2**
  - **Take context into account**
    - "Musashisakai" for me (when I spent 6 months at TUFS in 2009)
    - "Musashisakai" for you
    - "Musashisakai" for Tokyo inhabitants
  - **Short term memory**
    - "Musashisakai" = much simpler when TUFS has just been mentioned
    - $C(x_0) < \log_2(\#rank(e)|stm) + C(x_0 \mid e)$

Context-dependent relevant descriptions

- **Idea #3**
  - **Use contrast**
    - Entity → class → prototype → contrast → predicate → relevance
    - Musashino → place to live → Sth most desirable
    - Elvis → singer → n^{th} best-selling singer in history ($n = O(1)$)
    - Elvis → movie actor → N+ best-known films ($N >> 1$)

Context-dependent relevant descriptions

- **Problems**
  - **Retrieve entities**
  - **Retrieve predicates**
  - **Online computations of complexity**

Context-dependent relevant descriptions

- **Problems**
  - Retrieve entities
  - Retrieve predicates
  - Online computations of complexity
- **Good news**
  - No need to be exhaustive
  - No need to be efficent
  - DIG!!

INSTITUT POLYTECHNIQUE DE PARIS

TELECOM ParisTech

Merci pour votre attention

jean-louis @ dessalles.fr
www.dessalles.fr

Visit: www.simplicitytheory.science