

The YAGO-NAGA Approach to Knowledge Discovery

Kasneci et al

Apresentado por:
Rafael Odon de Alencar
odon@dcc.ufmg.br

Seminário de Gerência de Dados da Web - 17/06/2009

Roteiro

- 1. Extração e consulta de informação extensional**
- 2. A abordagem YAGO-NAGA**
 - 1. Visão Geral**
 - 2. Extraindo a base**
 - 3. Verificando a consistência**
 - 4. Crescendo/Atualizando a base**
 - 5. Consultando a base**
- 3. Conclusão**
- 4. Referências**

Extração e consulta de informação extensional

- Criar **bases de conhecimento universal** foi uma tentativa da comunidade **IA** por muitos anos.
- Como exemplos temos grandes **Ontologias e Tesouros** disponíveis:
 - *OpenCYC*,
 - *SUMO*,
 - *WordNet*,
 - *UMLS*.

Extração e consulta de informação extensional

- Essas bases estão focadas no conhecimento **intencional** das classes:
 - Descrição das classes num nível **conceitual**:
 - Atributos e seus domínios,
 - Restrições,
 - Relacionamentos com outras classes.
 - Ex:
*Matemático é uma Pessoa,
Pessoa pode ser Masculino ou Feminino,
Pessoa não pode voar.*

Extração e consulta de informação extensional

- Porém existe uma grande ausência de bases de conhecimento **extensional**:
 - Descrição de **instâncias** das classes descritas,
 - Relacionamentos **específicos** entre as instâncias.
- Ex: *Nicolas Sarkozy é o presidente da França.
França é um país.*

Extração e consulta de informação extensional

- A melhor fonte de conhecimento extensional é provavelmente a **Wikipedia**
 - Maior parte do conhecimento está embutido em textos **linguagem natural**,
 - Presença de alguns elementos semi-estruturados: **infoboxes** e sistemas de **categorias**.
- Além de conter tais fatos, uma boa base de conhecimento deveria permitir consultas.

Extração e consulta de informação extensional

- Aplicações possíveis:
 - **Enciclopédia** que pode ser lida por máquina,
 - **Busca semântica** na *Web*,
 - *Backbone* para **perguntas e respostas** em linguagem natural (Quem? Onde? Quando?),
 - **Tradução** automática e interpretação de textos (desambiguação na tradução),
 - Aquisição de conhecimento para **manter bases**.

A abordagem YAGO-NAGA:

Visão Geral

A abordagem YAGO-NAGA:

Visão Geral

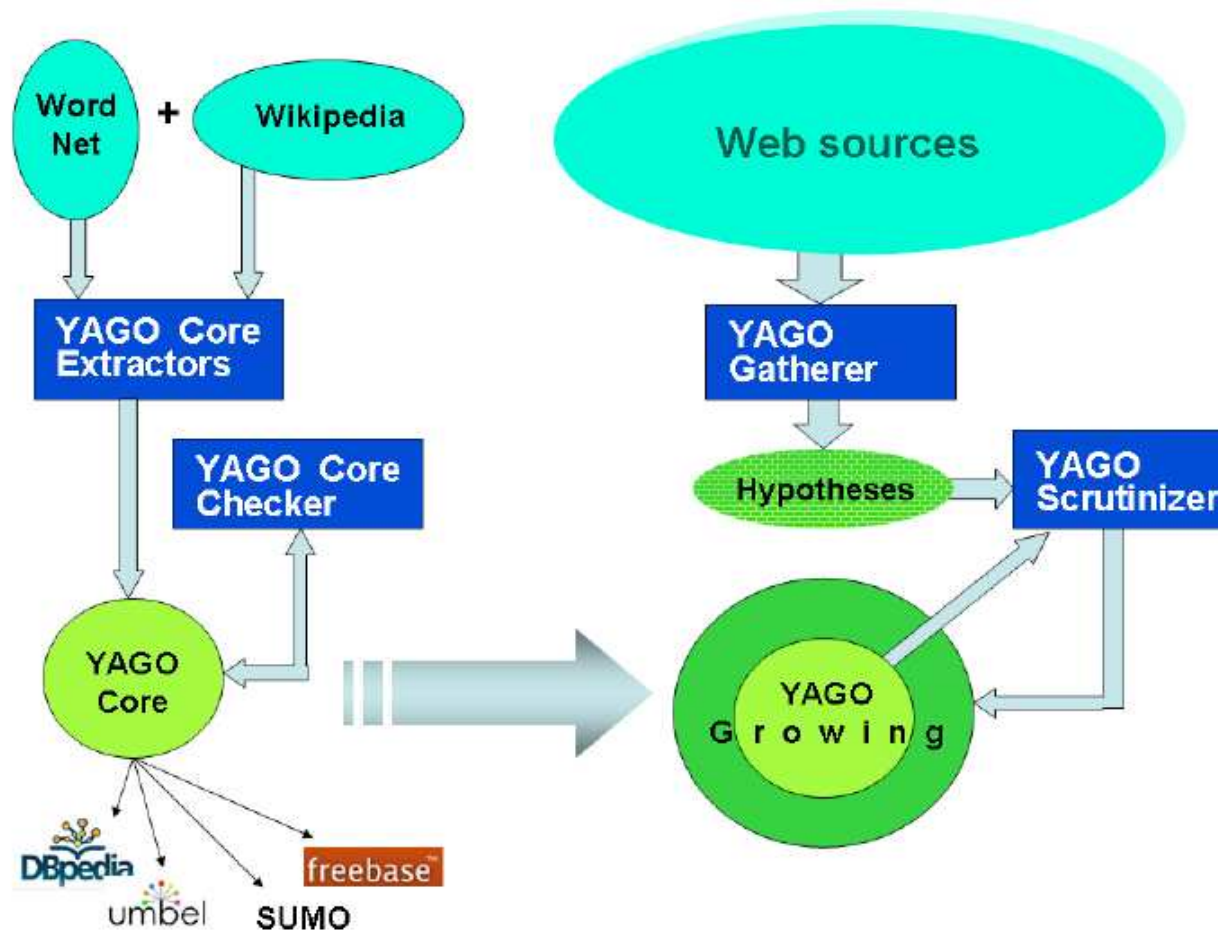
- Abordagem para **construir e manter** base de conhecimento:
 - consultável,
 - grande,
 - de alta precisão e consistência.
- YAGO (*Yet Another Great Ontology*)
 - Base + Técnicas de extração
 - Projeto que vem desde 2007
- NAGA (*Not Another Google Answer*)
 - Consulta + Ranking

A abordagem YAGO-NAGA:

Visão Geral

- Extrai uma **base de conhecimento inicial** de uma fonte *Web* tida como verdadeira e consistente.
- **Cresce/Atualiza** essa base com informações de fontes diversas.
- Possui mecanismos para checar **consistência** dos candidatos a fatos coletados.

A abordagem YAGO-NAGA: Visão Geral

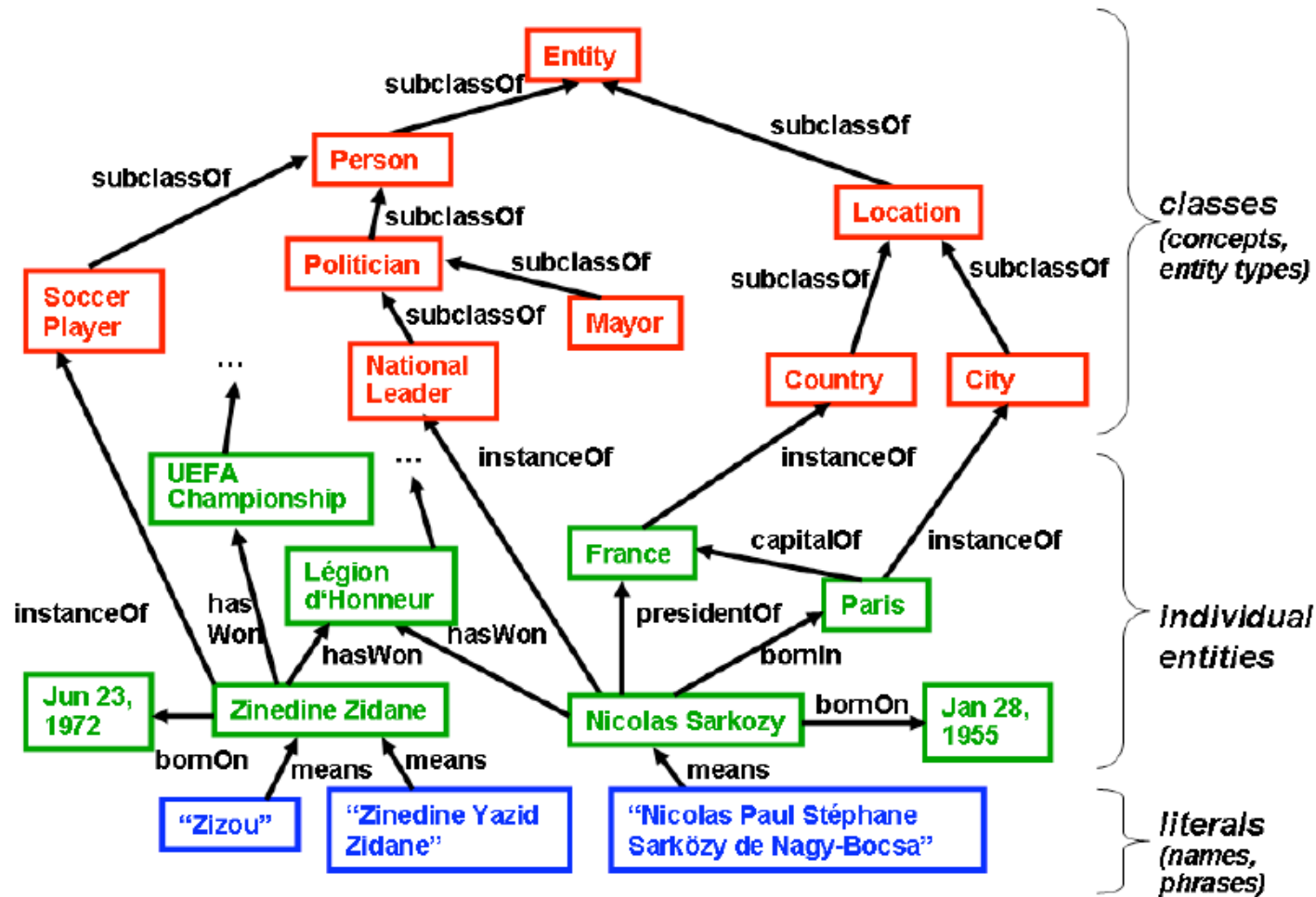


A abordagem YAGO-NAGA:

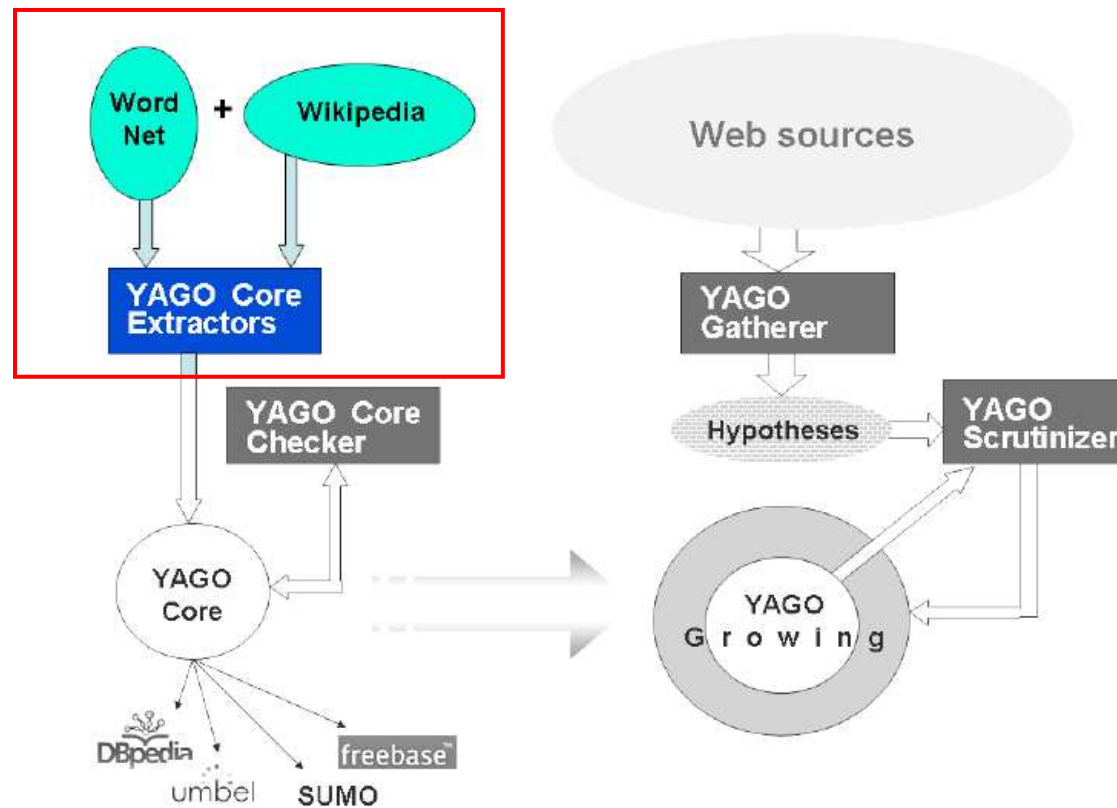
Visão Geral

- Armazena fatos como **grafo**, exportável para **RDF (Resource Description Framework)**:
 - Nós = entidades,
 - Arestas = relações.
 - Triplas RDF no formato: Sujeito, Predicado, Objeto.
 - *UFMG está-localizada-em Belo Horizonte*
- Permite **consultas** em linguagem baseada no **SPARQL** (Padrão W3C para consulta em RDF).

A abordagem YAGO-NAGA: Visão Geral



A abordagem YAGO-NAGA: Extraindo a base



A abordagem YAGO-NAGA: Extraíndo a base

- Prefere colher fatos precisos e consistentes a colher muitos fatos:
 - Baixa revocação, **alta precisão**.
- Executa a Extração de Informação (EI) nos artigos na *Wikipedia*:
 - Assume que a *Wikipedia* tem alta **informatividade e veracidade**.
- Checa consistência com as classes descritas no *WordNet*:
 - Precisão de pelo menos 95%.

A abordagem YAGO-NAGA: Extraindo a base

- Aproveita **infoboxes** e sistemas de **categorias** da *Wikipedia*.
- *Infoboxes*: no formato *atributo = valor*.
birth_date = 28 January 1955
birth_place = Paris
- **Mapeia** atributos para predicados: (?)
spouse => marriedTo
localização => locatedIn
- **Não tenta extrair tudo** de todos os *infoboxes*.

Domestic Dog



Labrador Retriever, a breed of domestic dog

Conservation status

Domesticated

Scientific classification

Kingdom: [Animalia](#)
Phylum: [Chordata](#)
Subphylum: [Vertebrata](#)
Class: [Mammalia](#)
Order: [Carnivora](#)
Family: [Canidae](#)
Genus: [Canis](#)
Species: [C. lupus](#)
Subspecies: [C. l. familiaris](#)

Trinomial name

Canis lupus familiaris

(Carl Linnaeus, 1758)

```
{{Infobox
| color = varied
| name = Domestic Dog
| status = DOM
| image = YellowLabradorLooking.jpg
| image_caption = [[Labrador Retriever]], a breed
of domestic dog
| regnum = [[Animal]]ia
| phylum = [[Chordate|Chordata]]
| subphylum = [[Vertebrate|Vertabrata]]
| classis = [[Mammal]]ia
| ordo = [[Carnivora]]
| familia = [[Canidae]]
| genus = ''[[Canis]]''
| species = ''[[Gray Wolf|C. lupus]]''
| subspecies = ''''C. l. familiaris''''
| trinomial = ''Canis lupus familiaris''
| trinomial_authority = ([[Carl Linnaeus]], 1758)
| category =
}}
```

A abordagem YAGO-NAGA: Extraíndo a base

- Aproveita sistema de **categorias** da *Wikipedia* para extrair relações do tipo *é-um*.
- Problema semântico:
 - Categoria *Calçada da Fama* possui vários artistas, no entanto esses artistas não são subclasses dessa categoria.
- YAGO tenta resolver com processamento lingüístico:
 - *Parsing* de **frases nominais** (frases sem verbo, ex: títulos)
 - Utilização de algumas **heurísticas**, ex: palavra chave da frase nominal deve estar no plural:
 - **Calçada** da Fama (não é superclasse)
 - **Cientistas** famosos (é superclasse).

[category](#)[discussion](#)[edit this page](#)[history](#)

Category:Portuguese-speaking countries

From Wikipedia, the free encyclopedia

The main article for this [category](#) is [Community of Portuguese Language Countries](#).

Pages in category "Portuguese-speaking countries"

The following 12 pages are in this category, out of 12 total. This list may not reflect recent changes ([learn more](#)).

*

- [Community of Portuguese Language Countries](#)

A

- [Angola](#)

B

- [Brazil](#)

C

- [Cape Verde](#)

E

- [East Timor](#)

G

- [Geographic distribution of Portuguese](#)
- [Guinea-Bissau](#)

M

- [Macau](#)

M cont.

- [Mozambique](#)

P

- [Portugal](#)
- [Flag of the Community of Portuguese Language Countries](#)

S

- [São Tomé and Príncipe](#)

[Categories: Countries by language](#) | [Foreign relations of Portugal](#) | [Portuguese language lists](#) | [Portuguese language](#)

A abordagem YAGO-NAGA:

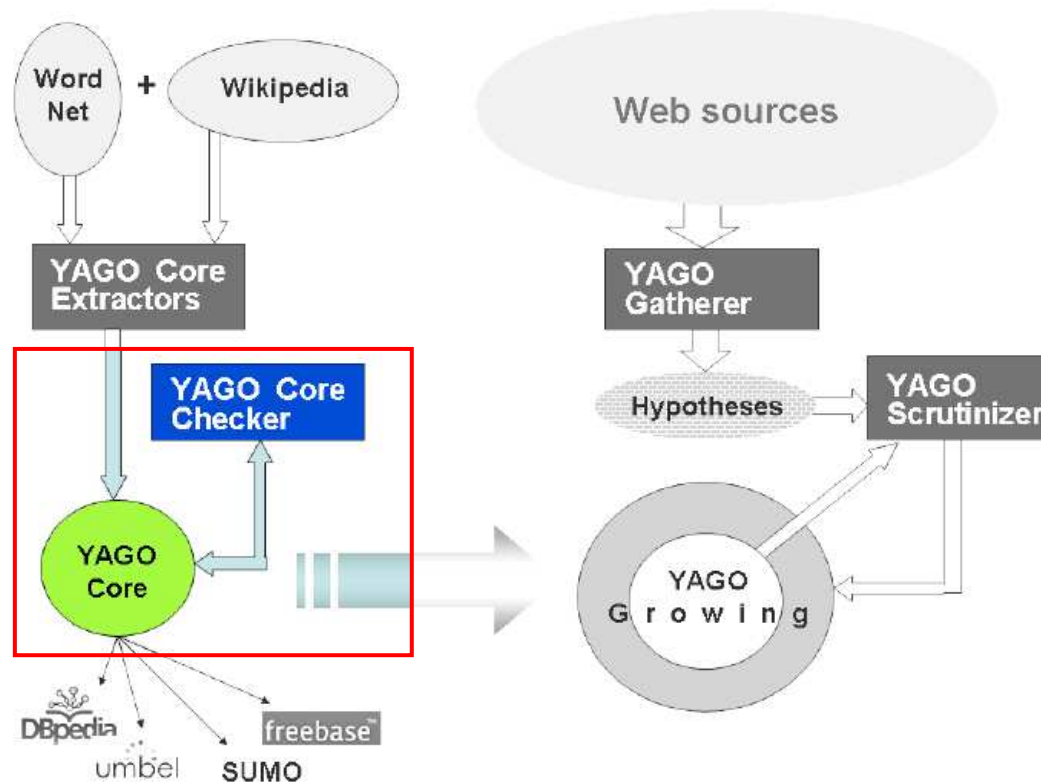
Extraindo a base

- Os fatos mudam com o **tempo** (Validade Temporal):
 - Ex: *Nicolas Sarkozy* é presidente hoje, mas não era 2 anos atrás.
- Propõe adaptação do **RDF**:
 - Adiciona ID's nas relações, criando outras:
id1 IsValidSince 17 May 2007.
- Estão trabalhando com extratores **orientados à texto** para detectar validade temporal.
 - Afirmações temporais podem referenciar pontos **relativos**. Ex:
Segunda-feira passada, No último Ano, Até a Copa.

A abordagem YAGO-NAGA: Extraíndo a base

- Tempo tem diferentes granularidades:
 - Ex: *Maio,*
Maio de 2007,
17 de Maio de 2007,
2007.
- YAGO representa cada ponto no tempo como um intervalo (*dataInicial - dataFinal*)
 - Permite refinar tempo dos fatos à medida que novas informações são coletadas.

A abordagem YAGO-NAGA: Verificando a consistência



A abordagem YAGO-NAGA: Verificando a consistência

- Extrações da *Wikipedia* podem resultar em uma coleção de dados potencialmente **inconsistentes**. Alguns dados não podem ser **inferidos** automaticamente. Ex:
Niocolas Sarkozy é presidente da França.
Logo, ele é um presidente.
Logo, ele faz parte do povo francês.
- Utiliza suporte taxonômico do *WordNet* para integrar e verificar consistência dos fatos.

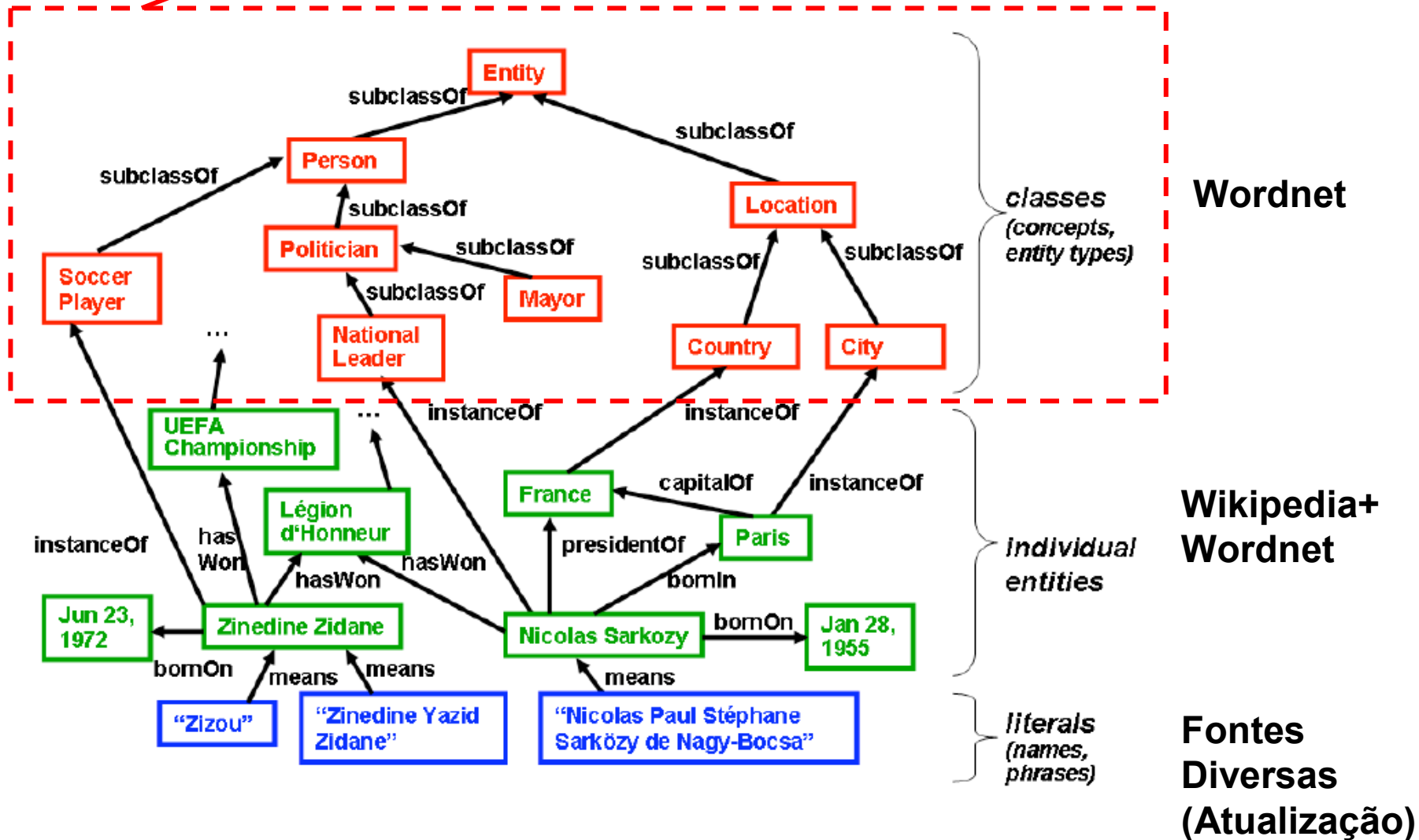
A abordagem YAGO-NAGA: Verificando a consistência

- Faz três verificações:
 - De hierarquia de classes
 - De tipos
 - De restrições nas relações (em andamento)

A abordagem YAGO-NAGA: Verificando a consistência

- Verificação de hierarquia de classes:
 - Cria sistema de classes **importando todas as classes da *WordNet*** com suas devidas relações de **subclasse/superclasse**.
 - Ao extrair informação da *Wikipedia*, se **não encaixar** no o sistema de classes, **descarta-a**.
 - Garante **alta consistência**.

Sistema de Classes Inicial



A abordagem YAGO-NAGA: Verificando a consistência

- Verificação de tipos:
 - Ex: define que a na relação *isCEOof* de um lado você tem uma Pessoa e de outro uma Empresa. (?)

Nicolas Sarkozy is CEO of France estaria incorreta pois França não é uma empresa.
 - Elimina afirmações ambíguas do tipo:

Pelé ama o Brasil.

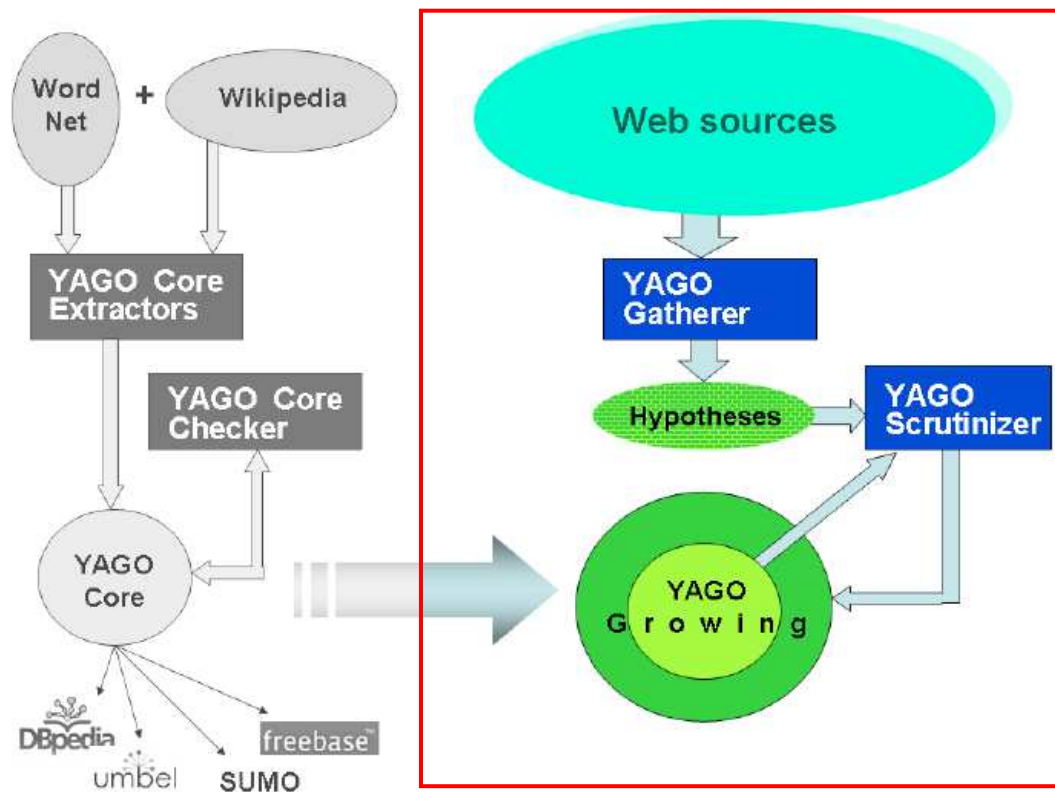
 - Pelé não é casado ou namorado do Brasil pois Brasil não é uma pessoa.

A abordagem YAGO-NAGA:

Verificando a consistência

- Verificação de restrições nas relações (em andamento):
 - Algumas relações da base de conhecimento podem podar **informações errôneas** da extração:
 - Uma pessoa só nasce em um local, logo, não coletar várias de nascimento.
 - Uma localidade não sofre alterações geralmente.
 - Algumas **inferências** podem ser feitas:
Nicolas Sarkozy é-presidente-da França,
Logo, França tem-como-presidente Nicolas Sarkozy.

A abordagem YAGO-NAGA: Crescendo/Atualizando a base



A abordagem YAGO-NAGA: Crescendo/Atualizando a base

- Necessário manter a base **atual, válida e útil**.
- Pode-se **rodar novamente** a extração na *Wikipedia + Wordnet*
 - Demora algumas horas,
 - Mantém os fatos já inferidos se não houver inconsistências.
- Pode-se extrair de outras fontes (*Gatherer*)
 - Textos, *websites*, *blogs*, artigos, notícias.
 - Utilização de técnicas de EI:
 - Casamento de Padrões,
 - Processamento de Linguagem Natural (PLN),
 - Aprendizado de Máquina.

A abordagem YAGO-NAGA: Crescendo/Atualizando a base

- Sugere o uso da ferramenta do grupo chamada LEILA:
 - **Heurísticas** para análise de **anáforas** (termos que referenciam outros já citados na sentença),
 - Produz grafo rotulado compatível com técnicas de aprendizado de máquina para classificar candidatos como **fatos** ou **hipóteses falsas**,
 - Boa precisão, mas demora alguns minutos para processar um artigo longo da *Wikipedia*.
- O YAGO permite utilizar várias ferramentas de PLN, mas os autores preferem não adentrar na escolha dessas técnicas.

A abordagem YAGO-NAGA: Crescendo/Atualizando a base

- As novas informações extraídas são **avaliadas** partindo do pressuposto que a base de conhecimento está correta.
- Inconsistências levam ao **descarte** de informações. (*Scrutinizer*)
- Pode-se aplicar as regras de verificação de consistência já descritas na fase de extração da base conhecimento.

A abordagem YAGO-NAGA: Crescendo/Atualizando a base

- Exemplos de **inconsistências**:
 - Uma pessoa só pode nascer em um único local,
 - Cidades européias não podem estar na América,
 - Americanos não podem nascer na Europa.
- Diversas outras inferências são possíveis.

A abordagem YAGO-NAGA: Crescendo/Atualizando a base

- Algumas classes **não estão bem representadas** na *Wikipedia*:
 - Nomes de medicamentos,
 - Cientistas da computação,
 - Entre outros.
- Fontes de conhecimento novas estão sendo adaptadas ao núcleo do YAGO como a **DBLP** e a **UMLS** (medicina).

A abordagem YAGO-NAGA: Consultando a base

A abordagem YAGO-NAGA:

Consultando a base

- Foi desenvolvida uma linguagem de consulta baseada na SPARQL (Padrão W3C para consulta em RDF)

SPARQL:

```
PREFIX abc: <http://example.com/ontology#>
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
      abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
      abc:isInContinent abc:Africa .
}
```

Versão NAGA:

```
$x isCapitalOf $y
$y isA Country
$y isInContinent Africa
```

A abordagem YAGO-NAGA:

Consultando a base

- Pode-se aplicar expressões regulares para flexibilizar a consulta e dar maior abrangência
- `$x (locatedIn)* Europe`
 - Qualquer cidade/região/estado da Europa
- `$x (citizenOf | bornIn | originatedFrom).(locatedIn)* Europe`
 - Qualquer entidade cidadã/nascida em/originada de alguma cidade/região/estado da Europa.

A abordagem YAGO-NAGA:

Consultando a base

- Consultas do tipo `connect()` (Em andamento)
 - `connect(Nicolas Sarkozy, Zinedine Zidane, Gerard Depardieu, Miles Davis)`
 - Resultado: todos foram membros da ordem nacional militar Legião da Honra Francesa.

A abordagem YAGO-NAGA: Consultando a base

- *Ranking:*
 - Utiliza uma técnica de **Modelo Estatístico de Linguagem** (LM) para capturar o grau de informatividade de um resultado
 - Presume que para uma pessoa é mais importante contemplar Nicolas Sarkozy nos resultados do que de um outro cidadão qualquer.

A abordagem YAGO-NAGA: Consultando a base

- *Ranking* :
 - Para isso propõe um modelo estatístico parecido com **TF-IDF**:
 - Verifica previamente em uma **amostra** as estatísticas de **ocorrências das classes**.
 - Resultados melhores são os que possuem mais **co-ocorrências** dos termos da consulta em sentenças, parágrafos ou páginas.
 - Ao procurar por ganhadores do *Grammy*, *Eric Clapton* e *Phill Collins* aparecerão primeiro por serem mais citados em textos sobre o prêmio.

A abordagem YAGO-NAGA:

Consultando a base

- Além das estatísticas de ocorrência, precisa associar a cada fato da base um **peso de confiança**:
 - Ocorre na fase de extração.
 - Considera quão boa foi a fonte de extração:
 - Alta confiança para *Wikipedia*,
 - Baixa confiança para *Blogs*.

A abordagem YAGO-NAGA:

Consultando a base

- Deseja-se que o comportamento usuário possa **personalizar** esses mecanismos de *ranking* (em andamento):
 - Monitora o **log de consultas** do usuário e os **cliques realizados**,
 - Usa os dados para alterar a relevância de alguns fatos.
- Se foi procurado muito sobre filmes épicos e música de orquestra, e pergunta-se sobre ganhadores do Grammy, seria priorizado Ennio Morricone entre outros criadores de trilhas de filmes, ao invés de Eric Clapton e Phill Collins.

Conclusão

Conclusão

- O trabalho é interessante mas contém muitas partes **em andamento** ainda.
- Alguns pontos não foram muito bem explicados pelos autores.
- Exige uma grande **intervenção humana**.
- Assume que a *Wikipedia* e outras bases são fontes de conhecimento correto e mais valioso que as demais fontes.
- Não deixa claro o **real ganho** do crescimento da YAGO por técnicas de EI em fontes **extra-*Wikipedia*** ou como isso altera a base existente.

Conclusão

- Correlação com a busca por palavras chave em XML:
 - NAGA utiliza de **técnicas estatísticas de RI** e para obter um bom **ranking** das respostas.
 - Base da YAGO é armazenada na forma de um **grafo**, representável por um RDF (**permite a busca** por palavra chave em XML).
 - Consulta via NAGA exige maiores inferências sobre o grafo, não apenas procurando **árvores mínimas de conexão entre termos** como no caso da busca por palavra chave.
 - Avaliar a busca semântica em XML, **mapeando palavras chaves e termos** do documento com a base de conhecimento YAGO.
 - Influência no *ranking* e na poda das respostas.

Referências

- Kasneci, G., Ramanath, M., Suchanek, F., and Weikum, G. 2009. *The YAGO-NAGA approach to knowledge discovery*. *SIGMOD Rec.* 37, 4 (Mar. 2009), 41-47.
- Weikum, G., Kasneci, G., Ramanath, M., and Suchanek, F. 2009. *Database and information-retrieval methods for knowledge discovery*. *Commun. ACM* 52, 4 (Apr. 2009), 56-64.

Fim

Perguntas?