Instituto de Computação

Relatório do Segundo Trabalho de Inteligência Artificial: Opção 1

Erick Grilo
Max Fratane
Vitor Lourenço
Vitor Santos

1 *Knowledge Bases*: Bases de conhecimento

1.1 Introdução

Uma base de conhecimento (do inglês, *knowledge base*, abreviadas KB) é uma espécie de repositório usada para armazenar dados complexos sobre os mais variados assuntos, que pode ser estruturados ou não, usados por sistemas computacionais. A grosso modo, uma base de conhecimento sobre um determinado assunto, ou sites, como a *Wikipedia* e até sessões de perguntas e respostas (*F.A.Q.*) podem ser considerads bases de conhecimento. Dentro do aspecto da computação uma KB é vista como um recurso que pode ser lido por máquinas com a intenção de disponibilizar informações. É usada em sistemas a fim de facilitar a coleção, organização e a extração de dados da mesma, por pessoas ou organizações.

O termo "base de conhecimento"foi cunhado em meados da década de 70, a fim de distinguir das bases de dados (bancos de dados). Nessa época, todas as informações eram armazenadas em bancos de dados hierárquicos ou relacionais, e a diferença entre as bases de dados era explícita: A base de dados possuía informações representadas como uma tabela, com *strings* ou números em cada elemento, usuários múltiplos, onde várias pessoas poderiam ter acesso à base de dados ao mesmo tempo, e dados estáticos e persistentes.

Os primeiros sistemas baseados em conhecimento necessitavam de dados que fossem armazenados de uma forma diferente de como os bancos de dados funcionam: necessitavam de ponteiros para outros objetos que, por sua vez, teriam ponteiros para outros dados. Tais sistemas visavam modelar o mundo que conhecemos, então, apenas dados armazenados em *strings* e inteiros não seriam suficientes. A representação ideal de conhecimento em bases de dados é com o uso de ontologia, que é a ideia de modelagem de um objeto, que pode possuir subclasses, interações com outras classes e instâncias.

Em um futuro (não tão distante), espera-se que humanos e bases de dados poderão trabalhar cooperativamente através da web semântica. A web semântica é uma extensão da web atual que procura padronizar a troca de dados e o formato de dados pela web usando o *Resouce Description Format* (RDF) (Miller 1998).A Web semântica interliga significados de palavras e, neste âmbito, tem como finalidade conseguir atribuir um significado (sentido) aos conteúdos publicados na Internet de modo que seja perceptível tanto pelo humano como pelo computador.

1.2 Alguns exmplos de *Knowledge Bases*

As bases de conhecimento atualmente possuem um papel crescente em melhorar a inteligência da Web, em buscas corporativas e em melhorar o suporte à integração da informação. A maioria de bases de conhecimento existentes cobrem apenas domínios específicos, são criados por grupos pequenos de pessoas e são extremamente custosas de se manter conforme os domínios variam. Nesse prisma, entram algumas bases de dados on-line especiais cujo foco é "aprender com a Web"e disponibilizar tais informações para todos. Alguns exemplos dessas bases são:

- YAGO (Yet Another Great Ontology): É uma base de conhecimento desenvolvida *Max Planck Institute for Computer Science*, em Saarbrücken, na Alemanha. YAGO é uma base de conhecimento semântica, que deriva (ou seja, absorve conhecimento) da Wikipedia, WordNet e do GeoNames. Atualmente, YAGO possui conhecimento de mais de 10 milhões de entidades (tais como pessoas, empresas, países, etc.). Algumas das especialidades do YAGO consiste em *YAGO, A High-Quality Knowledge Base* (n.d.):

  1. YAGO possui uma precisão de mais de 95% de veracidade de seus conhecimentos, que é avaliada manualmente.

  2. YAGO combina a taxonomia do *WordNet* University (2017) com a categorização da *Wikipedia*, atribuindo as entidades a mais de 350.000 classes.

  3. YAGO é uma ontologia que é âncorada no tempo e no espaço, ou seja, o YAGO atribui uma dimensão de tempo e uma dimensão de espaço para muitos de seus fatos e suas entidades. Além de uma taxonomia, YAGO possui domínios temáticos como "música", "pessoas", ou "ciência", provindos de domínios do *WordNet*.

  4. YAGO extrai e combina entidades e fatos de 10 Wikipedias de difentes idiomas.

  Os dados do YAGO são providenciados na sintaxe de *Turtle* (Terse RDF Triple Language), que é um formato para expressar dados em RDF, representando informações como triplas, que consiste em sujeito, predicado e objeto. Cada um desses itens é expressado como uma URI (Universal Resource Identifier) da Web. Seus dados também são expressados no formato *tsv* (*tab separated values*, um formato de texto simples que representa os dados em uma estrutura tabular.O YAGO também possui ligação com uma outra base de dados, a SUMO (http://www.site.uottawa.ca:4321/sumo/index.html)

- NELL (Never-Ending Language Learning) : É uma base de conhecimento desenvolvida na Universidade Carnegie Mellon, sob a alcunha de *Read the Web*, cujo principal objetivo é de ler e aprender com a Internet. Desde 2010, o NELL vem tentando fazer duas tarefas por dia: Primeiramente,

ele tenta ler (extrair) fatos de textos encontrados em milhões de páginas da web e, em seguida, procura melhorar sua habilidade de leitura, tal que futuramente ele consiga extrair mais fatos da Web, de forma mais precisa. o NELL possui mais de 50 milhões de crenças candidatas (a estarem corretas), que ficam divididas em diferentes níveis. Ele possui uma alta confidência em cerca de 3,565,291 dessas crenças.

Diferentes de outros sistemas, NELL aprende de forma autômata. Porém, como erros podem ocorrer, a cada duas semanas, a equipe do NELL procura por erros em sua base de dados e coloca o NELL para "ler a web"novamente.

Em Madureira et al. (2017), a NELL é citada como a primeira base de conhecimento que não para de operar. A base de dados do NELL é representada como uma estrutura baseada em ontologias, caracterizadas por categorias, relações e suas instâncias. Como a base de conhecimento do NELL expande todos os dias, ela não contém todas as instâncias de todas as categorais nem todas as instâncias de cada relação descritas na ontologia. No site do NELL, é possível fazer o download de dados específicos (obtidos em uma determinada iteração), a base de dados completa e a base de crenças candidatas. Tais dados são disponibilziados no fomato tsv (Tab-separated-value), onde a primeira linha determina os campos e as demais são os dados, separados por ".

- DBPedia : É uma base de dados que procura extrair informação estruturada da Wikipedia e fazer com que essa informação fique disponível na Web. Na DBPedia, é possível fazer "queries"em cima dos dados da Wikipedia, e ligar os diferentes conjuntos de dados da Web com os dados da Wikipedia, visando facilitar o trabalho de utilizar a informação já estruturada na Wikpedia para outros fins, e acabar por inspirar novos mecanismos para navegar, ligar e melhorar a própria Wikipedia.

  Como a Wikipedia é uma das maiores fontes de conhecimento da humanidade, mantida por milhares de colaboradores, o projeto da DBPedia procura extrair informações da Wikipedia e fazendo essas informações serem acessíves à todos na Web. Somente a versão da DBPedia em inglês descreve 4.58 milhões de coisas, onde desse número há 4.22 milhões classificadas em ontologias consistentes, incluindo pessoas, locais, músicas, jogos, espécies, companhias, etc.).

  Os dados na DBPedia são estruturados de acordo com os princípios de Tim-Berners Lee sobre dados ligados (cuja função, no contexto de Web Semântica, não é somente lançar os dados, mas também fazer com que a pessoa e a máquina possam explorar a web de dados.). Essa rede de dados ligados acaba resultando em bilhões de triplas em RDF, que cobrem domínios tais como pessoas, companhias, filmes e músicas, estruturada na forma de sujeito, predicado e objeto.

- Freebase : Freebase foi uma iniciativa de manter uma base de conhecimento composta principalmente pelos membros de sua comunidade. Era uma coleção de dados online estruturados e buscados de diversas fontes, incluindo contribuições individuais escritas em formato *wiki*. A ideia da Freebase era criar um recurso global que permitisse que máquinas e humanos acessassem informação comum de forma eficiente. Foi criado pela companhia Metaweb e funcionava publicamente desde março de 2007 adquirido pela Google em 2010. Foi desativado em maio de 2016. Deu origem ao *Knowlegde Graph*, da Google (2017).

  Freebase funcionava em uma base de dados estruturada pela própria Metaweb que utilizava um modelo de grafos, ao invés de usar tabelas e chaves como forma de definir as estruturas de dados. A Freebase definia sua estrutura de dados como um conjunto de nós e um conjunto de arestas que estabeleciam relações entre os nós do grafo. Como essa estrutura de dados não seguia uma hierarquia (como as encontradas em ontologias), era possível modelar estruturas de dados muito mais complexas entre indivíduos do que em bases de dados convencionais. As consultas eram feitas em uma linguagem chamada Metaweb Query Language, que era modelada sobre uma tripla chamada graphd.

2 O Trabalho

O trabalho consiste em extrair um fragmento de uma das bases de dados sugeridas e desenvolver uma busca que encontrasse um caminho entre duas entidades (dadas como entrada), na linguagem Prolog. A base de dados escolhida foi a NELL e as iterações escolhidas (os fragmentos) foram as iterações 333, 369, 666, 669, 690, 969 e a 999. A implementação pode ser encontrada em https://github.com/VitorA29/TrabalhoIA2-17.1

Dada uma interpretação errônea do trabalho, o grupo adotou duas metodologias de trabalho diferentes: uma baseada em consultas e outra em busca. Na primeira abordagem, foi desenvolvida uma ferramenta para formalizar a estrutura da *knowledge base* a partir do arquivo .csv dado na entrada. Em seguida, foram separadas todas as informações em quatro arquivos .pl: rneg.pl,que contém as negações das relações, fneg.pl, que contém a negação dos fatos, fpos.pl, contendo os fatos e rpos, contendo as relações. O arquivo consult-me.pl possui a ordem que tais arquivos devem ser consultados. Após a geração desses arquivos, foram desenvolvidas 10 consultas que se encontram no arquivo questions.pl

Já para a segunda abordagem, foi elaborada uma *query* de busca de relações entre entidades. Foi feita uma ferramenta que formaliza a *knowledge base*, onde a partir do .csv de entrada foi gerado o arquivo rpos.pl, formalizando a *knowledge base* em uma tupla de 3 informações. Tais informações foram definidas como (entidade,relação,valor). A busca foi elaborada baseada na ideia do caso base, que são entidades ligadas pela mesma relação ou elas estarem ligadas pelo mesmo valor. A busca então olha as ligações entre a primeira entidade e uma outra entidade que possem uma relação em comum ou um valor em comum, fazendo a ligação da nova entidade com a entidade de destino.

Referências

Google (2017), 'Google's Knowledge Graph'.
URL: *https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html*

Madureira, A. M., Abraham, A., Gamboa, D. & Novais, P. (2017), *Intelligent Systems Design and Applications: 16th International Conference on Intelligent Systems Design and Applications (ISDA 2016) held in Porto, Portugal, December 16-18, 2016*, Vol. 557, Springer.

Miller, E. (1998), 'An introduction to the resource description framework', *Bulletin of the American Society for Information Science and Technology* **25**(1), 15–19.

University, P. (2017), 'Wordnet,a lexical database for english'.
URL: *http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/*

YAGO, A High-Quality Knowledge Base (n.d.), http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/. Acessado em 08-06-2017.