Aprendizagem por reforço com redes neurais

Ricardo Yamamoto Abe

11 de junho de 2018

Objetivo

Dado um documento (ex. notícia), queremos encontrar documentos relacionados para oferecer ao usuário.
Hipótese de trabalho: apenas o conteúdo da coleção poderá ser utilizado no processamento.

Players

- Estadão – *tags* geradas pelos jornalistas.
- Google News – *clustering*.
- Wordpress (*plugins*)
  - *Related Entries* – busca pelos tokens do título.
  - *Similar Posts* – dispara um busca sobre os *n* tokens mais relevantes dentre título, conteúdo e *tags*.

Encontramos duas formas de resolver o problema:

- Construção de consultas baseadas no documento inicial.
- Cálculo de distância entre documentos e posterior agrupamento. (*Clustering*)

Construção de consultas

Dado um documento construímos uma consulta a ser realizada em um sistema de RI.
Dos resultados obtidos, supomos que os n primeiros documentos são relacionados ao documento original.

Construção de consultas

Temos duas opções:

- **Extrair termos** que sejam considerados relevantes ao documento inicial e utilizá-los em uma consulta a um sistema de RI tradicional (*mais à frente*).
- **Utilizar o próprio documento** como ponto de partida em um sistema baseado em CNG (*Contextual Network Graphs*)

Contextual Network Graphs

Contextual Network Graphs

Prós

- estrutura do grafo representa relacionamentos entre documentos
- baixo consumo de memória

Contras

- muitos parâmetros, não totalmente dominados
- algoritmo original pode consumir muito tempo
  - porém pode-se tentar reduzir as iterações
- seu uso em expansão de consulta não foi tão bem sucedido até agora

Cálculo de distância e agrupamentos

Nesse modelo, serão construídos agrupamentos de documentos (*clustering*) baseados em alguma métrica.
A partir de um documento inicial, serão escolhidos outros pertencentes ao mesmo agrupamento.

Clustering

Em um *clustering*, temos dois problemas principais:

- Cálculo de distância entre elementos.
- Metodologia de construção dos agrupamentos.

Distância

Existem basicamente dois tipos de distâncias entre documentos:

- Baseada em freqüência de termos.
- Baseada em semântica.

Distância baseada em freqüência

Para cada termo t num documento d é calculado um peso baseado na freqüência de t em d e na raridade de t na coleção.
Representamos cada documento como um vetor desses pesos e a distância entre dois documentos é a distância entre seus vetores.

Distância semântica

É calculada por meio de navegação em taxonomias, ou busca em dicionários.
Faz uso de bases como *WordNet* ou *Wikipedia*.

WordNet

A *WordNet* é um dicionário em que substantivos, verbos, adjetivos e advérbios são organizados em conjuntos de sinônimos, cada um representando um conceito, que são interligados por meio de relações semânticas.

Distância semântica

Usualmente, são utilizadas as seguintes métricas:

- Medida baseada em caminho: $f(length(c_1, c_2))$, onde *length* é o número de nós no menor caminho entre c_1 e c_2
- Medida baseada em conteúdo de informação: $g(max_{c \in S(c_1,c_2)}[-\log p(c)])$, onde $S(c_1, c_2)$ é o conceito que generaliza c_1 e c_2, e $p(c)$ é a probabilidade de encontrar o conceito c no *corpus*.
- Medida de sobreposição de texto: é uma medida do número de termos comuns entre as definições de dois conceitos.

Agrupamento em *Clustering*

Existem dois tipos:

- *Flat clustering* – cria os *clusters* sem uma estrutura explícita que relacione os *clusters* entre si.
- *Clustering* hierárquico – cria uma hierarquia, uma estrutura mais informativa que o conjunto não estruturado obtido no *flat clustering*.

Clustering

Prós

- É utilizado no Google News.
- Já foi utilizado no Data Quality.

Contras

- Depende da escolha de uma métrica e do agrupamento, o que aumenta o número de variáveis para otimização do processo.

Distância semântica

Prós

- Por usar relações semânticas entre conceitos, é capaz de trazer resultados mais relevantes.

Contras

- Definir os termos para comparação se análise for feita sobre todo o artigo. Em caso de existência de *tags*, é trivial.
- *WordNet* só existe em inglês e não cobre assuntos muito específicos.

Termos relevantes

Seja para elaborar uma consulta a um sistema de RI, ou para calcular distâncias, o uso de todos os termos presentes em um documento pode trazer problemas de desempenho ou até mesmo de qualidade das recomendações.

Termos relevantes

Algumas técnicas:

- Identificação de Sintagmas Nominais
- Identificação de Entidades Nomeadas
- Identificação de *keywords*

Termos relevantes

Algumas técnicas:

- **Identificação de Sintagmas Nominais**
- Identificação de Entidades Nomeadas
- Identificação de *keywords*

A bola de <SER HUMANO>Vitorino Ramos</SER> (na foto) é a materialização de um mapa cognitivo de um formigueiro artificial .

Termos relevantes

Algumas técnicas:

- Identificação de Sintagmas Nominais
- **Identificação de Entidades Nomeadas**
- Identificação de *keywords*

A { [bola] de [Vitorino] [Ramos] } (em { a [foto] }) é { a [materialização] de um [mapa] cognitivo de um [formigueiro] artificial } .

Termos relevantes

Algumas técnicas:

- Identificação de Sintagmas Nominais
- Identificação de Entidades Nomeadas
- **Identificação de keywords**

A bola de [Vitorino] Ramos (na foto) é a [materialização] de um mapa [cognitivo] de um [formigueiro] artificial .

Identificação de Sintagmas Nominais

Faz uso de ferramentas de *POS tagging* e TBL para identificação de Sintagmas Nominais em textos.

Identificação de Sintagmas Nominais

Prós

- Já foi utilizado aqui na upLexis (para o CLEF)
- Aplicação de SN em expansão de consulta traz bons resultados
- Os sintagmas (ou pelo menos os substantivos) representam boa parte da informação de um documento

Contras

- Faz uso de aprendizado de máquina, logo podem haver erros de identificação
- Realiza marcações no texto (pode ser necessário manter duas versões de cada documento)

Identificação de Entidades Nomeadas

A “Linguateca” desenvolveu uma ferramenta denominada SIEMÊS para isso, além de um repositório de entidades nomeadas classificadas, o REPENTINO.

Identificação de Entidades Nomeadas

Prós

- As entidades nomeadas que ocorrem em um documento tendem a ser a parte mais importante do mesmo.

Contras

- Envolve regras fixas e uma base de categorias em constante manutenção, portanto podem haver erros de identificação e de classificação
- Realiza marcações no texto (pode ser necessário manter duas versões de cada documento)

Identificação de *keywords*

Podemos utilizar *tf-idf* e escolher os n primeiros termos como os mais relevantes do documento.

Termos relevantes

Podemos também utilizar *tags* definidas pelos usuários ou pelo autor para representar um documento.
Tags podem ser usadas como medida de distância entre documentos, calculada pela quantidade de sobreposição de *tags*.

LSA

Geralmente, uma coleção de documentos é tratada como sendo um espaço determinado por uma matriz Termos-por-Documentos, onde cada elemento corresponde ao peso de um termo em um documento.

LSA

Latent Semantic Analysis (LSA ou Análise Semântica Latente) é uma forma de decompor essa matriz Termos-por-Documentos, permitindo a redução da dimensionalidade desta, mas preservando o máximo possível de informação da matriz original.

LSA

A decomposição obtida (normalmente através de *Singular Value Decomposition*), consiste em 3 matrizes; a primeira define um espaço Termos-por-Conceitos; a segunda define um "peso" para cada conceito; e a terceira define um espaço Conceitos-por-Documentos.
Uma forma de identificar documentos recomendados é, dado o documento original, determinar quais estão mais próximos a ele no espaço Conceitos-por-Documentos.

LSA

Prós

- Diminui a dimensão do espaço a ser trabalhado
- Permite determinar similaridade entre documentos em termos de “conceitos”

Contras

- Matrizes envolvidas deixam de ser esparsas, e portanto consomem mais memória
- Cálculo consome muito processamento
- Difícil de implementar

Fim