A User-Centered Concept Mining System for Query and Document Understanding at Tencent

分享人：Sm1les

论文创新点

构建了一套以用户为中心的标签（概念）体系

现存方法的缺点

1. 纯模板的方法召回有限
2. 人工标注过于主观
3. 传统方法只能抽连续的词
4. 当前方法时效性差

Methodology

STEP-1

概念挖掘

Methodology

Bootstrapping by Pattern-Concept Duality

数据：query

方法：

- 用模板去抽概念
  - eg：十大XXX
- 用概念反推模板
  - eg：哪款XXX性能好?

Methodology

Concept mining by query-title alignment
数据：query + top N clicked title
方法：用query去和top N clicked title进行对齐匹配
eg：

- query：
  香港僵尸电影
- top 2 clicked title:
  1.香港最后一部僵尸电影
  2.香港搞笑僵尸电影
- 对齐匹配出概念：
  香港僵尸电影

Methodology

Supervised sequence labeling
数据：带有明显语法标志的句子
方法：序列标注
eg：

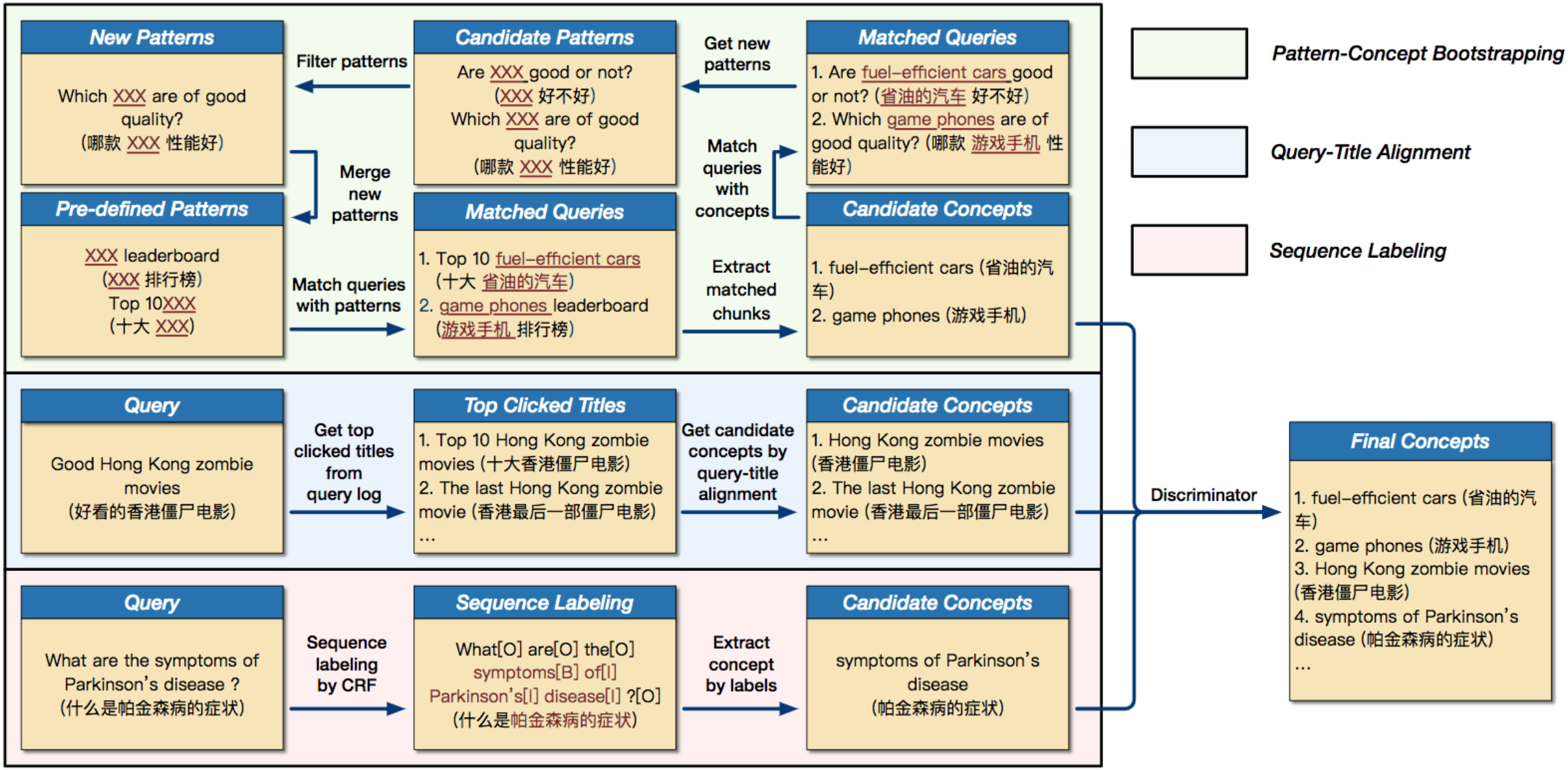
- data:什么 是 帕金森病 的 症状
- label:O O B I I

Methodology

A Discriminator for quality control

数据：通过以上方法挖掘出来的概念

方法：训练一个分类器来判断挖掘出来的概念的质量


New Patterns
Which XXX are of good quality?
(哪款 XXX 性能好)
Filter patterns
Candidate Patterns
Are XXX good or not?
(XXX 好不好)
Which XXX are of good quality?
(哪款 XXX 性能好)
Get new patterns
Matched Queries
1. Are fuel-efficient cars good or not? (省油的汽车 好不好)
2. Which game phones are of good quality? (哪款 游戏手机 性能好)
Merge new patterns
Match queries with concepts
Pre-defined Patterns
XXX leaderboard
(XXX 排行榜)
Top 10XXX
(十大 XXX)
Match queries with patterns
Matched Queries
1. Top 10 fuel-efficient cars
(十大 省油的汽车)
2. game phones leaderboard
(游戏手机 排行榜)
Extract matched chunks
Candidate Concepts
1. fuel-efficient cars (省油的汽车)
2. game phones (游戏手机)
Query
Good Hong Kong zombie movies
(好看的香港僵尸电影)
Get top clicked titles from query log
Top Clicked Titles
1. Top 10 Hong Kong zombie movies (十大香港僵尸电影)
2. The last Hong Kong zombie movie (香港最后一部僵尸电影)
...
Get candidate concepts by query-title alignment
Candidate Concepts
1. Hong Kong zombie movies (香港僵尸电影)
2. The last Hong Kong zombie movie (香港最后一部僵尸电影)
...
Query
What are the symptoms of Parkinson's disease ?
(什么是帕金森病的症状)
Sequence labeling by CRF
Sequence Labeling
What[O] are[O] the[O] symptoms[B] of[I] Parkinson's[I] disease[I] ?[O]
(什么是帕金森病的症状)
Extract concept by labels
Candidate Concepts
symptoms of Parkinson's disease
(帕金森病的症状)
Discriminator
Final Concepts
1. fuel-efficient cars (省油的汽车)
2. game phones (游戏手机)
3. Hong Kong zombie movies (香港僵尸电影)
4. symptoms of Parkinson's disease (帕金森病的症状)
...
Pattern-Concept Bootstrapping
Query-Title Alignment
Sequence Labeling

Methodology

STEP-1

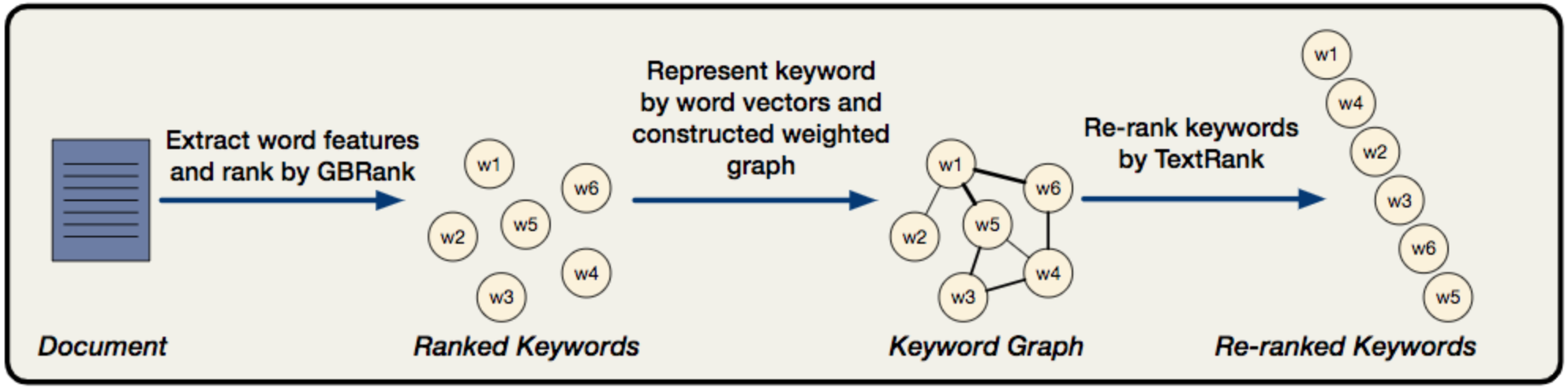
概念-文档匹配

Methodology

Key instance extraction

数据：文档

方法：关键词抽取

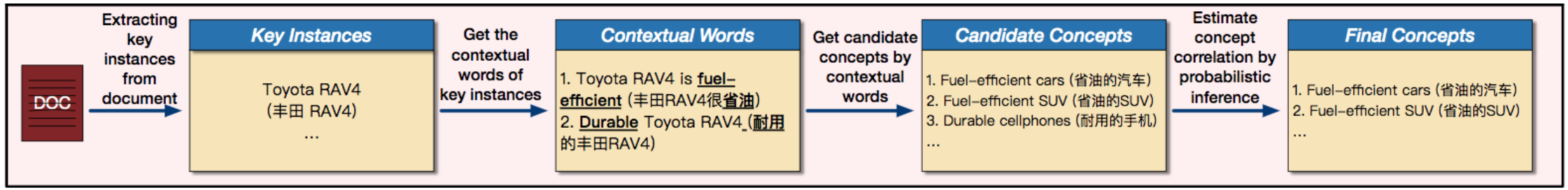


Methodology

Concept tagging by probabilistic inference

数据：文档 + key instance

方法：根据key instance的上下文去反推概念



Methodology

Concept tagging by matching

数据：文档 + key instance + 标签（概念）体系（Taxonomy）

方法：直接根据key instance的isA(son of)关系直接匹配