LIA at TREC 2011 Web Track

Online resources combination and Wikipedia thematic graphs

Romain Deveaud[α], Eric SanJuan[α] and Patrice Bellot[β]

[α] LIA – University of Avignon
[β] LSIS – Aix-Marseille University

Introduction

- 2 main ideas experimented this year
- Wikipedia contextual expansion
- Online resource combination
- All query expansion related
- Showed to be effective on book search at INEX 2010

Introduction (cont.)

- Building a Wikipedia **thematic** and **query-driven** graph
  - Expansion terms
  - Anchor texts
  - Linking topic-related articles together
- Combining information from 2 online resources
  - Google web search
  - Wikipedia

Outline

- Introduction
- Online resources combination
- Wikipedia thematic graphs
- Conclusions

Online resources combination

- Using external sources of information for web search
- Using multiple sources of information [Diaz2006]
- Query expansion with each resource taken separately…
- … but how about combining them?

Online resource combination (cont.)

- Two online resources queried on July, 2011:
  - Wikipedia API
  - Google API
- Wikipedia English contained by the ClueWeb
- Google « contains » the ClueWeb (to a certain extent)

Online resource combination (cont.)

- Using the API of each resource to retrieve their **first ranked** document
- Topic 111: lymphoma in dogs
  - Wikipedia best ranked article: http://en.wikipedia.org/wiki/Lymphoma_in_animals
  - Google best ranked page: http://www.caninecancer.com/Lymphoma.html
- Computing an *entropy* measure for each word w of a Wikipedia article $\mathcal{W}$:

$$H_{\mathcal{W}}(w) = -\sum_{w \in \mathcal{W}} p_{\mathcal{W}}(w) \cdot \log p_{\mathcal{W}}(w)$$

Online resource combination (cont.)

- Selecting the top 20 words
- Expanding the query with the selected words, using their *entropy* to weigh them inside the query
- Example of words extracted from the article *Lymphoma in Animals*:

```
0.14528344114469402  lymphoma
0.07659549574716934  cats
0.05273200888401322  dogs
0.04716238190543428  cell
0.041381311223068006 treatment
0.039401174024736065 lymph
0.039401174024736065 chemotherapy
0.03739176127955678  veterinary
0.03535111847579512  proceedings
0.03535111847579512  common
0.03535111847579512  disease
0.03535111847579512  symptoms
0.03327701121637184  loss
0.031166860398936912 nodes
0.031166860398936912 gastrointestinal
0.029017655680289543 prognosis
0.026825837286266783 remission
…
```

Online resource combination (cont.)

- Non-parametric LM approach to retrieval
- Each resource counts as a single word in the expansion

```
#combine (
    #weight ( 0.85 #combine ( lymphoma in dogs )
              0.1  #combine ( #1(lymphoma in) #1(in dogs) )
              0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )
    #weight ( 0.14528344114469402 lymphoma
              0.07659549574716934 cats
              0.05273200888401322 dogs … )
    #weight ( 0.08508212158684461 cancer
              0.04398871871995759 bone … )
          )
```

Online resource combination (cont.)

- Non-parametric LM approach to retrieval
- Each resource counts as a single word in the expansion

```
#combine (
    #weight ( 0.85 #combine ( lymphoma in dogs )
               0.1  #combine ( #1(lymphoma in) #1(in dogs) )
               0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )
    #weight ( 0.14528344114469402 lymphoma
               0.07659549574716934 cats
               0.05273200888401322 dogs … )
    #weight ( 0.08508212158684461 cancer
               0.04398871871995759 bone … )
          )
```

Sequential Dependance Model [Metzler2005]

Online resource combination (cont.)

- Non-parametric LM approach to retrieval
- Each resource counts as a single word in the expansion

```
#combine (
    #weight ( 0.85 #combine ( lymphoma in dogs )
               0.1  #combine ( #1(lymphoma in) #1(in dogs) )
               0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )
    #weight ( 0.14528344114469402 lymphoma
               0.07659549574716934 cats
               0.05273200888401322 dogs … )
    #weight ( 0.08508212158684461 cancer
               0.04398871871995759 bone … )
          )
```

Sequential Dependance Model [Metzler2005]

Wikipedia

Online resource combination (cont.)

- Non-parametric LM approach to retrieval
- Each resource counts as a single word in the expansion

```
#combine (
    #weight ( 0.85 #combine ( lymphoma in dogs )
               0.1  #combine ( #1(lymphoma in) #1(in dogs) )
               0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )
    #weight ( 0.14528344114469402 lymphoma
               0.07659549574716934 cats
               0.05273200888401322 dogs … )
    #weight ( 0.08508212158684461 cancer
               0.04398871871995759 bone … )
          )
```

Sequential Dependance Model [Metzler2005]

Wikipedia

Google

Online resource combination (cont.)

- Indri for indexing and searching
- Krovetz stemming and stopwords removal
- No spam filter
  - How does the LM behave on the **entire web**?
  - No weight optimization, entire **non-parametric** system
- 2 submitted runs on category A, 1 run on category B

Online resource combination (cont.)

Run	Resources	MAP	P@20	nDCG@20	ERR@20
SDM (unofficial)	-	0.1111	0.1270	0.0963	0.0409
liaQEWikiA	**Wiki**	0.1323**	0.2500***	0.1567**	0.0519**
GooA (unofficial)	**Google**	0.1438***	0.2140***	0.1868***	0.0825***
liaQEWikiGoA	**Wiki + Google**	**0.1566*****	0.2780***	0.1978***	0.0765***
liaQEGoo	**Wiki + Google (cat B)**	0.1321	**0.3260**	**0.2228*****	**0.0952*****
NoSpamSDM	-	**0.1651**	0.3370	0.2390	0.1216
NoSpamWikiGoA	**Wiki + Google**	0.1628	**0.3600**	**0.2635**	**0.1230**

Table 1: Comparison of the retrieval performance of the four submitted runs and two additional runs. We use two sided paired wise Wilcoxon test ($^{*}: p < 0.1$; $^{**}: p < 0.05$; $^{***}: p < 0.01$) to determine significant differences with baseline.

- Each resource improves baseline on its own
- Combining the resources achieves best MAP
- The ClueWeb09 category B is a « purest » sub-collection

Outline

- Introduction
- Online resources combination
- Wikipedia thematic graphs
- Conclusions

Wikipedia thematic graphs

- Several approaches build a complete Wikipedia graph *a priori*

  [Milne2008, Coursey2009, Yeh2009]

- No contextual information

- Building a **query-oriented** Wikipedia graph could highlight more thematic relations between articles

Wikipedia thematic graphs

- Given a query (topic 111) :

  lymphoma in dogs

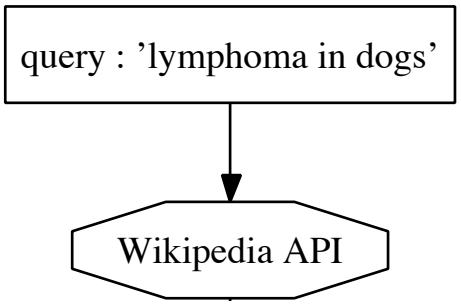
query : 'lymphoma in dogs'

Wikipedia thematic graphs

- Given a query (topic 111) :

  `lymphoma in dogs`

- Query Wikipedia API

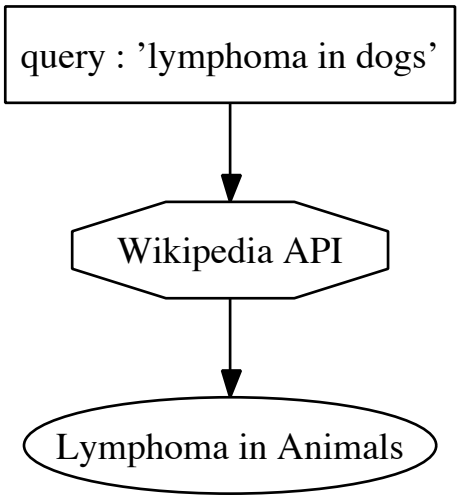


Wikipedia thematic graphs

- Given a query (topic 111) :

  `lymphoma in dogs`

- Query Wikipedia API
- Retrieve first article



Wikipedia thematic graphs

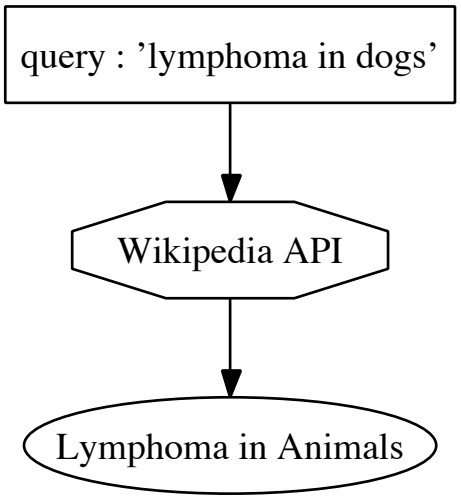
- Given a query (topic 111) :

  `lymphoma in dogs`

- Query Wikipedia API

- Retrieve first article

- Compute *entropy* of each word and select top 20




```
[
    lymphoma, dogs, cell, treatment, lymph,
    chemotherapy, gastrointestinal, prognosis…
]
```

Wikipedia thematic graphs (cont.)

- Compute *entropy* of each word and select top 20

[

lymphoma, dogs, **cell**, treatment, **lymph**, **chemotherapy**, **gastrointestinal**, **prognosis**…

]

Wikipedia thematic graphs (cont.)

- Compute *entropy* of each word and select top 20

- Intersect anchor texts with selected words and select top 2

[
lymphoma, dogs, **cell**, treatment, **lymph**, **chemotherapy**, **gastrointestinal**, **prognosis**…
]

[
lymph node,
white blood **cell**,
gastrointestinal,
remission…
]

Wikipedia thematic graphs (cont.)

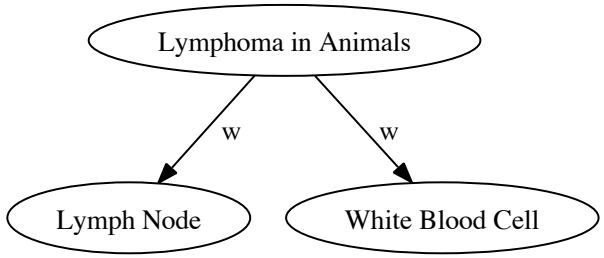
- Compute *entropy* of each word and select top 20

```
[
    lymphoma, dogs, cell, treatment, lymph,
    chemotherapy, gastrointestinal,
    prognosis…
]
```

- Intersect anchor texts with selected words and select top k

```
[
lymph node,
white blood cell,
gastrointestinal,
remission…
]
```

- Instantiate graph with the first article and two linked *sub-articles*



Wikipedia thematic graphs (cont.)

- Expansion terms are extracted from all nodes of the graph
- Same term extraction method as before
- Terms from *sub-articles* are weighted half (w = 0.5)

Wikipedia thematic graphs (cont.)

- Expansion terms are extracted from all nodes of the graph
- Same term extraction method as before
- Terms from *sub-articles* are weighted half (w = 0.5)

```
#weight (
    1.0 #weight ( 0.85 #combine ( lymphoma in dogs )
                  0.1  #combine ( #1(lymphoma in) #1(in dogs) )
                  0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )

    1.0 #weight ( 0.14528344114469402 lymphoma
                  0.07659549574716934 cats
                  0.05273200888401322 dogs … )
    0.5 #combine (
                   #weight ( 0.18625434418721107 lymph
                             0.09701593966960126 nodes … )
                   #weight ( 0.09773065837466471 prognosis
                             0.05711842409792383 disease … )
                )
        )
```

Wikipedia thematic graphs (cont.)

- Expansion terms are extracted from all nodes of the graph
- Same term extraction method as before
- Terms from *sub-articles* are weighted half (w = 0.5)

```
#weight (
   1.0 #weight ( 0.85 #combine ( lymphoma in dogs )
                 0.1  #combine ( #1(lymphoma in) #1(in dogs) )
                 0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )

   1.0 #weight ( 0.14528344114469402 lymphoma
                 0.07659549574716934 cats
                 0.05273200888401322 dogs … )
   0.5 #combine (
                  #weight ( 0.18625434418721107 lymph
                            0.09701593966960126 nodes … )
                  #weight ( 0.09773065837466471 prognosis
                            0.05711842409792383 disease … )
               )
       )
```

Sequential Dependance Model [Metzler2005]

Wikipedia thematic graphs (cont.)

- Expansion terms are extracted from all nodes of the graph
- Same term extraction method as before
- Terms from *sub-articles* are weighted half (w = 0.5)

```
#weight (
    1.0 #weight ( 0.85 #combine ( lymphoma in dogs )
                  0.1  #combine ( #1(lymphoma in) #1(in dogs) )
                  0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )

    1.0 #weight ( 0.14528344114469402 lymphoma
                  0.07659549574716934 cats
                  0.05273200888401322 dogs … )
    0.5 #combine (
                    #weight ( 0.18625434418721107 lymph
                              0.09701593966960126 nodes … )
                    #weight ( 0.09773065837466471 prognosis
                              0.05711842409792383 disease … )
                )
        )
```

Sequential Dependance Model [Metzler2005]

Lymphoma in Animals

Wikipedia thematic graphs (cont.)

- Expansion terms are extracted from all nodes of the graph
- Same term extraction method as before
- Terms from *sub-articles* are weighted half (w = 0.5)

```
#weight (
    1.0 #weight ( 0.85 #combine ( lymphoma in dogs )
                  0.1  #combine ( #1(lymphoma in) #1(in dogs) )
                  0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )

    1.0 #weight ( 0.14528344114469402 lymphoma
                  0.07659549574716934 cats
                  0.05273200888401322 dogs … )
    0.5 #combine (
                    #weight ( 0.18625434418721107 lymph
                              0.09701593966960126 nodes … )
                    #weight ( 0.09773065837466471 prognosis
                              0.05711842409792383 disease … )
                )
        )
```

Sequential Dependance Model [Metzler2005]

Lymphoma in Animals

Lymph Node

Wikipedia thematic graphs (cont.)

- Expansion terms are extracted from all nodes of the graph
- Same term extraction method as before
- Terms from *sub-articles* are weighted half (w = 0.5)

```
#weight (
    1.0 #weight ( 0.85 #combine ( lymphoma in dogs )
                  0.1  #combine ( #1(lymphoma in) #1(in dogs) )
                  0.05 #combine ( #uw8(lymphoma in) #uw8(in dogs) ) )

    1.0 #weight ( 0.14528344114469402 lymphoma
                  0.07659549574716934 cats
                  0.05273200888401322 dogs … )
    0.5 #combine (
                    #weight ( 0.18625434418721107 lymph
                              0.09701593966960126 nodes … )
                    #weight ( 0.09773065837466471 prognosis
                              0.05711842409792383 disease … )
                )
        )
```

Sequential Dependance Model [Metzler2005]

Lymphoma in Animals

Lymph Node

White Blood Cell

Wikipedia thematic graphs (cont.)

Run	Resources	MAP	P@20	nDCG@20	ERR@20
SDM (unofficial)	-	0.1111	0.1270	0.0963	0.0409
liaQEWikiA	**Wiki**	**0.1323****	0.2500***	0.1567**	0.0519**
liaQEWikiAnA	**Wiki graph**	0.1218	**0.2610*****	**0.1630****	**0.0606****
NoSpamWikiAnA	**Wiki graph**	0.1259	0.3110	0.2167	0.1008

Table 2: Comparison of the retrieval performance of the Wikipedia thematic graph approach and simple Wikipedia expansion. We use two sided paired wise Wilcoxon test (* : $p < 0.1$; ** : $p < 0.05$; *** : $p < 0.01$) to determine significant differences with baseline.

- Thematic graphs improves early precision
- Also performs better for graded metrics
- Slight loss in average precision

Outline

- Introduction
- Online resources combination
- Wikipedia thematic graphs
- Conclusions

Conclusions

- Experiments with the use of two online resources
- Generation of a thematic graph for expanding the query
- Combining the resources largely improves the results, especially early precision

Ongoing work

- Reproducing results with offline resources
  - System with on-board indexes
  - Using the non-spammed ClueWeb09 as a clean web resource

Run	Resources	MAP	P@20	nDCG@20	ERR@20
SDM	-	0.1111	0.1270	0.0963	0.0409
indexWikiA	**Wiki index**	0.1095	**0.2600**	**0.1737**	**0.0766**
liaQEWikiA	**Wiki API**	**0.1323**	0.2500	0.1567	0.0519

Table 3: Comparison of query expansion using the Wikipedia API or an index of a Wikipedia dump. Retrieval is performed on the entire ClueWeb09 without spam-filtering.

Thank you for your attention

romain.deveaud@univ-avignon.fr