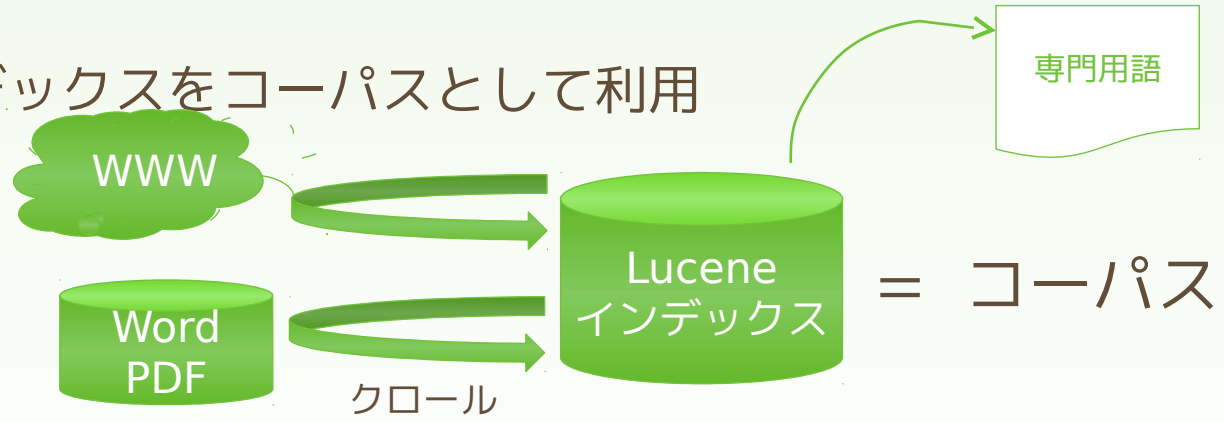
Luceneインデックスからの専門用語抽出

関口宏司@ロンウイット

背景／目的

- 専門用語抽出
  - サジェスチョン
  - クラスタリング
  - MLT
  - もしかして検索

- Luceneインデックスをコーパスとして利用



専門用語抽出

- 単名詞
- 複合名詞
- ターム性
  - ある言語単位（複合名詞など）の持つ専門分野固有の概念への関連性の強さ
  - 専門文書を書いた専門家の概念に直結していると考えられる
- ユニット性
  - ある言語単位がコーパス中で安定して使用される度合い

手順

単名詞バイグラムと頻度

単名詞Nの右方に連接する単名詞　　RNiの頻度(i=1…m)

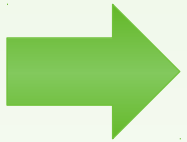
[LN1　N](#L1)　　[N　RN1](#R1)
[LN2　N](#L2)　　[N　RN2](#R2)
　：　　：
[LNn　N](#Ln)　　[N　RNm](#Rm)

単名詞Nの左方に連接する単名詞　　LNiの頻度(i=1…n)

連接種類数　　n → #LDN(N)
m → #RDN(N)

実際の例（頻度と異なり数）

トライグラム　統計
トライグラム
単語　トライグラム
クラス　トライグラム
単語　トライグラム
トライグラム
トライグラム　抽出
単語　トライグラム　統計
トライグラム
文字　トライグラム



[単語　トライグラム](3)
[クラス　トライグラム](1)
[文字　トライグラム](1)

#LDN(トライグラム)=3

[トライグラム　統計](2)
[トライグラム　抽出](1)

#RDN(トライグラム)=2

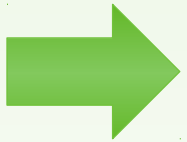
連接頻度

$$\sharp LN(N) = \sum_{i=1}^{n} (\sharp L_i)$$

$$\sharp RN(N) = \sum_{i=1}^{m} (\sharp R_i)$$

実際の例（連接頻度）

トライグラム　統計
トライグラム
単語　トライグラム
クラス　トライグラム
単語　トライグラム
トライグラム
トライグラム　抽出
単語　トライグラム　統計
トライグラム
文字　トライグラム



[単語　トライグラム](3)
[クラス　トライグラム](1)
[文字　トライグラム](1)

#LDN(トライグラム)=3
#LN(トライグラム)=5

[トライグラム　統計](2)
[トライグラム　抽出](1)

#RDN(トライグラム)=2
#RN(トライグラム)=3

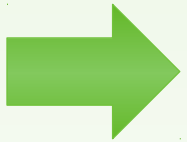
複合名詞のスコアリング

$$LR(CN) = \left\{ \prod_{i=1}^{L} (FL(N_i) + 1)(FR(N_i) + 1) \right\}^{\frac{1}{2L}}$$

単名詞：Ni
複合名詞：CN=N1N2...NL
単名詞Nの左方スコア関数：FL(N)
単名詞Nの右方スコア関数：FR(N)

実際の例（CNスコア、連接頻度の場合）

トライグラム　統計
トライグラム
単語　トライグラム
クラス　トライグラム
単語　トライグラム
トライグラム
トライグラム　抽出
単語　トライグラム　統計
トライグラム
文字　トライグラム



[単語　トライグラム](3)
[クラス　トライグラム](1)
[文字　トライグラム](1)

#LDN(トライグラム)=3
#LN(トライグラム)=5

[トライグラム　統計](2)
[トライグラム　抽出](1)

#RDN(トライグラム)=2
#RN(トライグラム)=3

$$LR(CN) = \left\{ \prod_{i=1}^{L} (FL(N_i)+1)(FR(N_i)+1) \right\}^{\frac{1}{2L}}$$

$$LR(\text{トライグラム}) = \sqrt{(5+1)(3+1)} \approx 4.90$$

出現頻度を考慮した重み付け

$$FLR(CN) = f(CN) \times LR(CN)$$

$$FLR(\text{トライグラム}) = 3 \times \sqrt{(5+1)(3+1)} \approx 14.70$$

C-value

$$C_{value}(CN) = (length(CN) - 1) \times \left\{ n(CN) - \frac{t(CN)}{c(CN)} \right\}$$

CN: 複合名詞
length(CN): CNの長さ（名詞数）
n(CN): コーパスにおけるCNの出現回数
t(CN): CNを含むより長い複合名詞の出現回数
c(CN): CNを含むより長い複合名詞の異なり数

MC-value

$$MC_{value}(CN) = length(CN) \times \left\{ n(CN) - \frac{t(CN)}{c(CN)} \right\}$$

$$MC_{value}(\text{トライグラム}) = 1 \times \left\{ 10 - \frac{7}{6} \right\} \approx 8.83$$

※ 参考文献[1]とカウント結果が異なる

CN: 複合名詞
length(CN): CNの長さ（名詞数）
n(CN): コーパスにおけるCNの出現回数
t(CN): CNを含むより長い複合名詞の出現回数
c(CN): CNを含むより長い複合名詞の異なり数

実際の抽出例

※ 連接頻度FLR法（Lucene近似版）を利用

- livedoorニュースコーパス（以下同様）、スポーツウォッチ（上位120位まで）

日本	世界	関係	女子サッカー	一日	勝利
選手	プロ	サッカーファン	注目	なでしこ	代表選手
日本代表	ファン	女子	移籍	巨人	発表
試合	同番組	プロ野球選手	開催	決勝	北海道日本ハムファイターズ
放送	ロンドン五輪	日本中	リーグ	人気	決勝戦
サッカー	試合後	ゴール	野村氏	最終的	田中将大
野球	解説	自分	試合前	選手たち	サッカー界
代表	スポーツ番組	野村克也氏	ロンドン五輪出場	代表戦	番組中
監督	優勝	出演	五輪出場	世界選手権	日本シリーズ
チーム	野球解説者	本田圭佑	獲得	前出	一回
番組	サッカー女子日本代表	ネット上	掲示板上	女子W杯	五輪代表
五輪	元日本代表	関係者	自身	所属	最終戦
プロ野球	試合中	日本戦	日本人選手	日本サッカー	日本代表監督
W杯	プロ野球解説者	プレー	インタビュー	オープン戦	予選
一戦	週刊アサヒ芸能	関連リンク	二人	問題	ツイッター上
一人	相手	ブログ	開幕	コメント	本田
番組内	サッカー選手	スポーツ	ボール	生出演	開幕戦
サッカー解説者	大会	練習	Jリーグ	一年	可能性
サッカー日本代表	なでしこジャパン	報道	球団	日本サッカー協会	北京五輪
出場	野球選手	楽天	スポーツ選手	日本選手権	スポーツ紙

実際の抽出例

※ 連接頻度FLR法（Lucene近似版）を利用

- トピックニュース（上位140位まで）

ネット掲示板	同番組	livedoorニュース	可能性	二人	芸能	独島
関連情報	女性	関係	人気	記事内	さん	韓国名
関連記事	コメント	番組中	選手	タレント	放送後	発売
記事	韓国人	ファン	生活	人たち	社会	政府
ネットユーザー	番組内	写真ギャラリー	ネットユーザ	女性自身	記事中	フジテレビ系
ネット	視聴者	メディア	AKB	芸人	子供	記者
ネット上	日韓	ユーザ	五輪	各メディア	ドラマ	日本中
番組	掲示板上	livedoor	テレビ番組	女子	外国人	関連リンク
日本	韓国語	関係者	ネットユーザーたち	韓国ニュース	日本人	同番組内
放送	発言	韓国メディア	事件	逮捕	活動	代表
批判	ツイッター	ツイッター上	フジテレビ	公開	騒動	大阪
ユーザー	批判的	韓国ネット掲示板	韓国内	橋下氏	映画	ラジオ番組
韓国	写真	発表	自分	生活保護	被害	同ブログ
ネット掲示板上	情報番組	公式サイト	視聴率	怒り	事故	CM
同記事	テレビ	韓国ネットユーザー	自身	コメント欄	ツイート	番組放送中
ニュース	出演	被害者	男性	否定的	アイドル	共演者
情報	一人	ニュースサイト	メンバー	内容	自殺	芸能人
報道	韓流	放送中	話題	問題視	結婚	韓国人女性
問題	意見	芸能界	出演者	関連	事務所	日本代表
掲示板	ブログ	放送事故	殺到	紹介	世界	日韓戦

スコア計算方法の違いによる抽出比較

連接種類数LR法	連接頻度LR法	連接種類数FLR法	連接頻度FLR法
選手	日本	選手	日本
日本中	日本代表	日本	選手
日本選手	代表	日本代表	日本代表
日本戦	日本代表選手	チーム	試合
女子選手	日本選手	試合	放送
代表選手	代表選手	監督	サッカー
日本代表選手	日本代表戦	サッカー	野球
日本	サッカー日本代表	一人	代表
日本代表戦	日本戦	一戦	監督
代表戦	日本女子代表	野球	チーム
女子代表選手	女子日本代表	代表	番組
選手ら	選手	ファン	五輪
一日	日本中	出場	プロ野球
一日一日	女子野球日本代表	番組	W杯
同選手	代表戦	優勝	一戦
リー選手	五輪代表	世界	一人
中一	日本サッカー	番組内	番組内
一選手	日本代表監督	同番組	サッカー解説者
サッカー選手	野球選手	試合後	サッカー日本代表
野球選手	女子代表選手	プレー	出場

※ FLR法はいずれもLuceneによる近似実装

参考文献

[1] 出現頻度と連接頻度に基づく専門用語抽出

中川、湯本、森 自然言語処理 Vol. 10 No. 1 Jan 2003

[2] Lucene 4.0 Javadoc

http://lucene.apache.org/core/4_0_0/index.html

[3] livedoor ニュースコーパス

http://www.rondhuit.com/download.html#ldcc