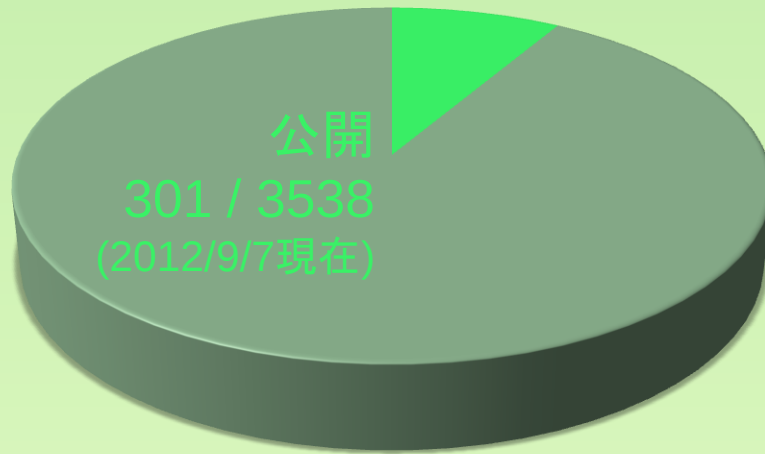


ASKサービス事例の テキストマイニング

2013年3月6日

参考調査係 岩井

埋もれた事例を発掘する



ASK事例があまり
活用されていない

テキストマイニングによって、
使える事例が掘り起こせるのでは？

「**KH Coder**」の使用

- テキストマイニングツール
<http://khc.sourceforge.net/>
- 抽出から統計分析までワンストップ
－茶筌、Rが組み込まれている

分析対象

- 対象としたASK事例
 - 2008年3月31日までに完結したもの
(2013年4月に削除予定) **1052件**
 - 質問文と回答文のみ抽出
- 形態素解析を行ったのち、品詞による取捨選択を行った
 - 形容詞や感動詞等は除外

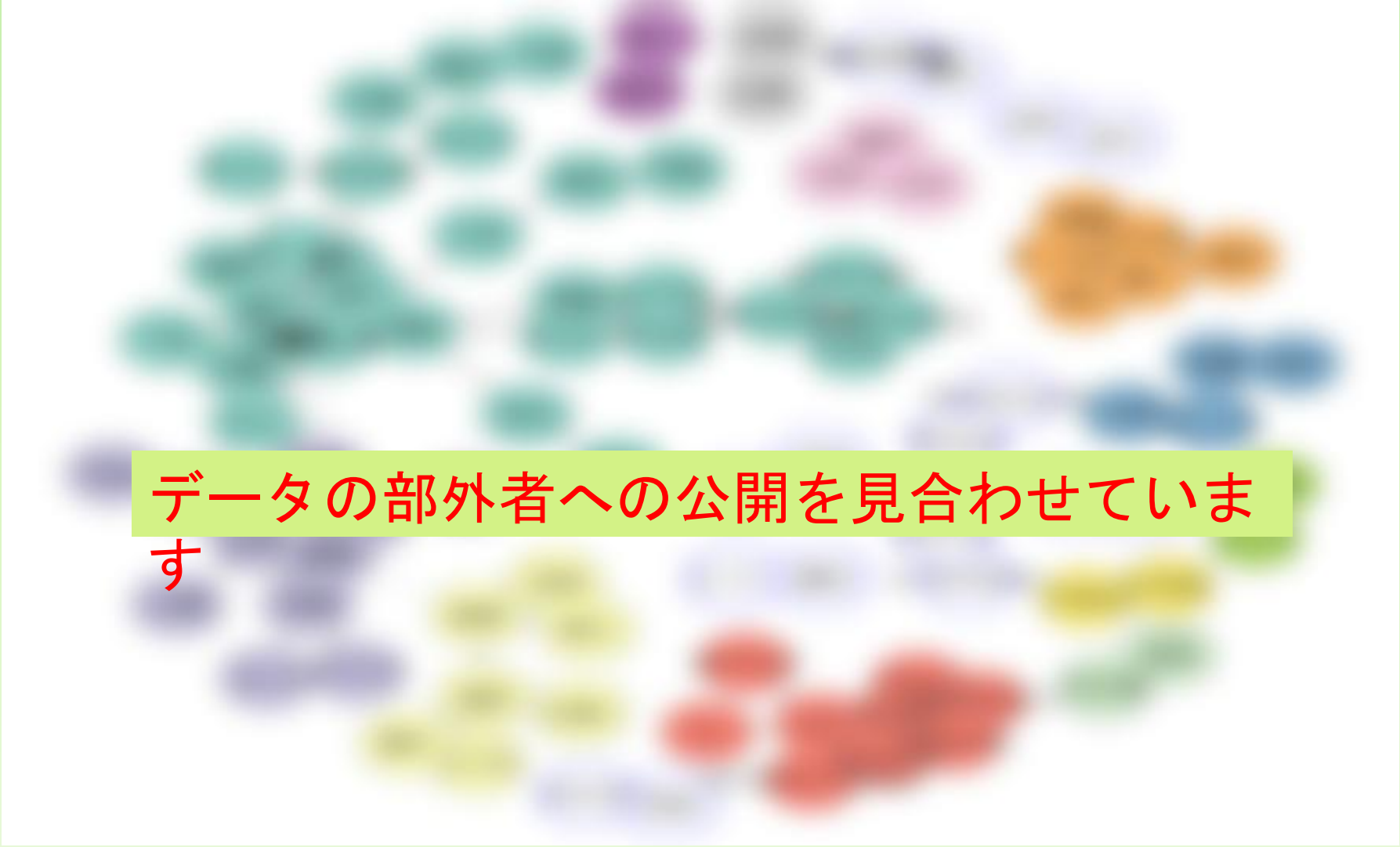
品詞別の出現文書数（抜粋）

[illegible]

使用した解析機能

- 共起ネットワーク
 - 同じ文書で同時に現れる頻度の高い語
- 自己組織化マップの作成
 - 語の類似関係
- 文書クラスタ分析
 - 含まれる語により文書をクラスタ分け

共起ネットワーク



データの部外者への公開を見合わせています

自己組織化マップ



データの部外者への公開を見合わせています

文書クラスタ分析

各クラスタの特徴語

所蔵調査...？

農学系...？

電子ジャー
ナルにつな
がらない系
...？

データの部外者への公開を見合わせていま

す

データベー
スの使い方
とか...？

英語関係！

オンライン
関係っぽい
...？

分析結果をどうみるか

- 今回の閾値は20文書以上出現、200語強を対象に分析している
→単語分析に現れた語は全て頻出語
- よくある質問パターンが見えてくる
- クラスタ分けされた事例を個別に見ていくと、頻出パターンがより分かりやすい

解析対象とする語彙の問題

- 同義語の処理などは全くおこなっていない
- 固有名詞などは回数が少ないので大半が対象外に
- 閾値を下げると解析時間膨大化

今後の応用は？

- 同等の機能を、たとえばWebアプリ化する...？
 - －システム面もデータ面も課題が多そう