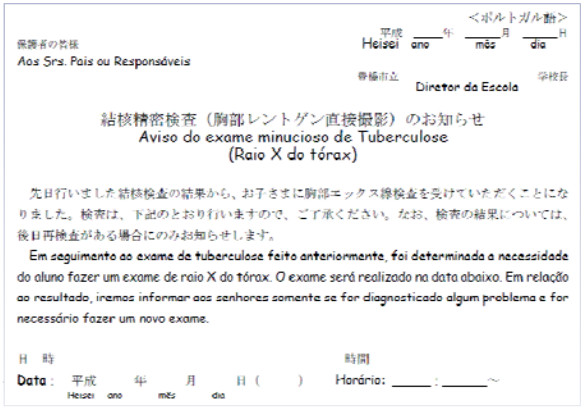
多言語校務文書ポータルサイトにおけるキーワードサジェスト機能の実装とユーザ評価

澤晃平*1 中山晋也*2 堀雅洋*1 喜多千草*1
*1：関西大学大学院総合情報学研究科　*2：関西大学総合情報学部

多言語校務文書とは

- 日本語を母語としない児童生徒のための校務文書
  - 日本語と外国語を併記
  - 教育委員会関連サイトなどで公開
- 内容
  - 行事
  - 保健
  - 届出・証明書
  - ・・・
- ファイルタイプ
  - PDF，Word など



＜ポルトガル語＞

保護者の皆様
Aos Srs. Pais ou Responsáveis

平成 Heisei ＿＿年 ano ＿＿月 mês ＿＿日 dia

＿＿市立 ＿＿学校長
Diretor da Escola

結核精密検査（胸部レントゲン直接撮影）のお知らせ
Aviso do exame minucioso de Tuberculose
(Raio X do tórax)

先日行いました結核検査の結果から、お子さまに胸部エックス線検査を受けていただくことになりました。検査は、下記のとおり行いますので、ご了承ください。なお、検査の結果については、後日再検査がある場合にのみお知らせします。

Em seguimento ao exame de tuberculose feito anteriormente, foi determinada a necessidade do aluno fazer um exame de raio X do tórax. O exame será realizado na data abaixo. Em relação ao resultado, iremos informar aos senhores somente se for diagnosticado algum problema e for necessário fazer um novo exame.

日時　　時間
Data：平成 Heisei ＿＿年 ano ＿＿月 mês ＿＿日 dia（　）　Horário: ＿＿：＿＿～

多言語校務文書ポータルサイトとは

- 全国で公開されている多言語校務文書を分類・整理し，リンク集としてURLを集約

ポータルサイト利用者動向

- 2009年10月より一般公開
  - http://www.tagengo-gakko.jp/bunsho/
  - 毎月の平均ユニークユーザ数 478名（2010年1月～12月）
- 文書検索状況（対象期間: 2010年1月～12月）
  - キーワードを入力しダウンロード：471件
    - 有効なキーワード例：「成績証明書」「授業参観」
  - キーワードを入力したがダウンロードなし：216件
    - 有効でないキーワード例：
      - 語数の少ない一般的な語：「健康」「連絡」
      - 意味が限定される複合語：「マラソン前健康診断」「放課後子ども教室」

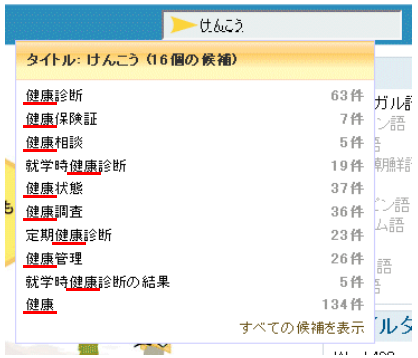
⇒ キーワードの想起を手助けするなど，ユーザ補助システムの導入が必要

本日の発表

1. キーワード検索支援システムの概要
   - キーワードサジェスト機能
   - 領域知識（主題分類を用いた索引語の重み付け）
2. 前回の評価実験
   - サジェスト機能と領域辞書の有効性検証[澤 2010]
3. 実ユーザの協力を得て実施した評価実験
   - ユーザの前提知識が検索成功率に与える影響を検証
4. サジェスト機能とファセット検索を併用した評価実験
   - 多言語校務文書検索以外の領域で評価を実施

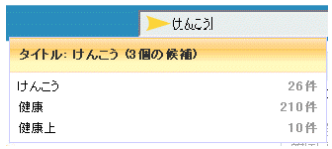
キーワードサジェスト機能

- 入力されたキーワードと部分一致する索引語を候補として表示
- 索引語は，形態素解析によって各文書に付加

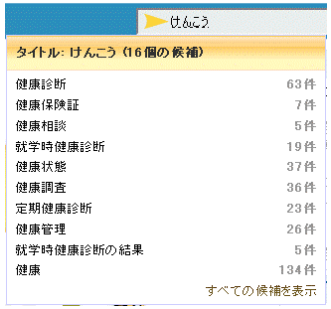


領域辞書による索引語抽出

- 形態素解析の際，一般的な語を領域固有の語に拡張
  - 「健康」⇒「健康診断」「健康保険証」…
- 領域辞書として教育分野関連用語集(予定表語彙など)約2700語を利用



(a) 領域辞書を用いない場合



(b) 領域辞書を用いた場合

索引語の重み付け：従来方法での問題点

- 従来手法　TF-IDF法
  - 索引語の出現頻度に基づく重み付け方式
  - 本サイトに適用すると頻繁に登場する索引語（「検査」など）が大きく重み付けられる
    - IDF値による高頻度単語の重み引き下げの効果が薄い
    - 本サイトでは有望でない

⇒ 領域知識に基づく重み付け方式の提案

「け」に対するサジェスト結果

順位	TF-IDF法 索引語	提案手法 索引語
1	検査	保険証
2	健康	保健体育
3	検診	携帯電話
4	結果	鍵盤ハーモニカ
5	健康診断	視力検査
6	結核	保健室
7	保健	聴力検査
8	付け	入学試験
9	保険	歯科検診
10	見学	内科検診
…	…	…

領域知識に基づく索引語の重み付け [澤 2010]

- 本手法の方針
  - (1)領域辞書に出現する索引語は価値が高い
  - (2)多くの主題に均等に出現する索引語は価値が低い
    - 主題:「行事」「保健」「届出・証明書」など9種類
    - 平均情報量として算出
  - (3)ひらがな2文字で構成された索引語は価値が低い

(1)辞書にあれば定数X_iを加算

(2)平均情報量Y_iを加算

(3)ひらがな2文字なら半減

$$W_i=(X_i+Y_i)\times Z_i/2$$

キーワードサジェスト機能のデモ

評価方針

- 文書検索における領域辞書およびサジェスト機能の効果測定
- 領域辞書を用いたサジェスト機能の背景知識を持ったユーザに対する効果測定
  - 背景知識を用いたキーワードの厳選の可能性
- 実験要因
  - サジェスト機能(あり, なし)
  - 領域辞書(あり, なし)
  - ユーザ種別(学生, 実ユーザ)

実験計画

- ■ 領域辞書 (あり, なし) [参加者内要因] } 前回報告
- ■ サジェスト機能 (あり, なし) [参加者間要因] } 今回の実験
- ■ ユーザ種別 (学生, 実ユーザ) [参加者間要因]

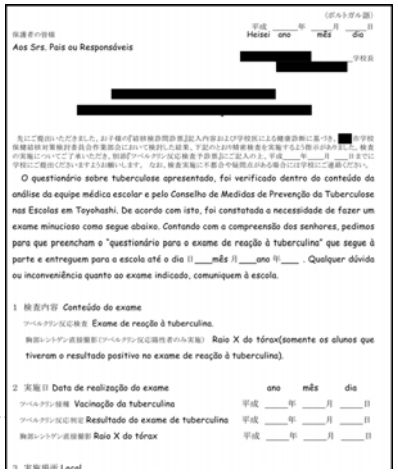
ユーザ種別グループ

サジェスト機能ありグループA		サジェスト機能なしグループB	
A1	A2	B1	B2
領域辞書あり(2課題)	領域辞書なし(2課題)	領域辞書あり(2課題)	領域辞書なし(2課題)
↓	↓	↓	↓
領域辞書なし(2課題)	領域辞書あり(2課題)	領域辞書なし(2課題)	領域辞書あり(2課題)

(課題の実施順序は各グループ内で相殺)

前回の評価実験 [澤 2010]

- 目的
  - サジェスト機能と領域辞書による検索成功率向上の検証
- 実験協力者
  - 情報系学部生32名
    （うち女性9名）
- 検索対象
  - ポルトガル語文書（1028件）
- 課題
  - 探索目的と正解文書例を提示
  - 全部で4課題
    - 制限時間は1課題5分間


Aos Srs. Pais ou Responsáveis

O questionário sobre tuberculose apresentado, foi verificado dentro do conteúdo da análise da equipe médica escolar e pelo Conselho de Medidas de Prevenção da Tuberculose nas Escolas em Toyohashi. De acordo com isto, foi constatada a necessidade de fazer um exame minucioso como segue abaixo. Contando com a compreensão dos senhores, pedimos para que preencham o "questionário para o exame de reação à tuberculina" que segue à parte e entreguem para a escola até o dia 日 ___mês 月 ___ano 年___ . Qualquer dúvida ou inconveniência quanto ao exame indicado, comuniquem à escola.

1 検査内容 Conteúdo do exame

ツベルクリン反応検査 Exame de reação à tuberculina.

胸部レントゲン直接撮影（ツベルクリン反応陽性者のみ実施） Raio X do tórax(somente os alunos que tiveram o resultado positivo no exame de reação à tuberculina).

2 実施日 Data de realização do exame ano mês dia

ツベルクリン接種 Vacinação da tuberculina 平成 ____年 ____月 ____日

ツベルクリン反応判定 Resultado do exame de tuberculina 平成 ____年 ____月 ____日

胸部レントゲン直接撮影 Raio X do tórax 平成 ____年 ____月 ____日

前回の実験結果：検索成功率

2要因（2×2）の分散分析

	1分未満	2分未満	3分未満
領域辞書 （参加者内要因）	* 辞書なし＜あり 0.16　0.30	* 辞書なし＜あり 0.49　0.66	---
サジェスト機能 （参加者間要因）	---	---	* 機能なし＜あり 0.70　0.83

* $p<0.05$

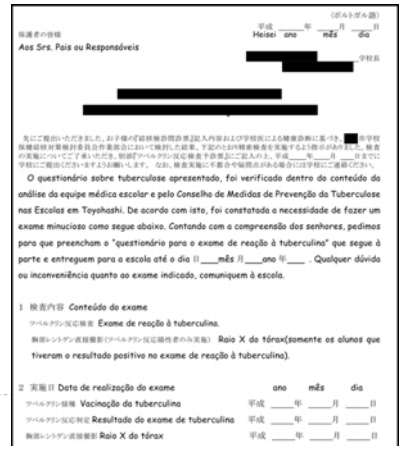
- 1,2,3分未満全ての場合で，交互作用なし（$p > 0.05$）
- 3分までででほとんどの協力者が実験を完了

前回の実験結果：まとめ

- 2分未満で領域辞書利用による検索成功率が上昇
  - 探索時間が短いうちは領域辞書によって特徴を強めた索引語による検索支援が有効
- 2～3分ではサジェスト機能の利用による検索成功率が上昇
  - 探索困難な文書については，サジェスト機能によってキーワードの想起を助けることが有用

実ユーザによる
多言語校務文書ポータルサイト上での評価実験

- 目的
  - サジェスト機能と領域辞書による検索成功率向上の検証
  - ユーザ種別が検索成功率に与える影響の検証
- 実験協力者
  - 多言語教育環境に関わりをもつ教員・翻訳者など24名
    （うち女性19名）
- 検索対象
  - ポルトガル語文書（1028件）
- 課題
  - 探索目的と正解文書例を提示
  - 全部で4課題
    - 制限時間は1課題5分間


Aos Srs. Pais ou Responsáveis

O questionário sobre tuberculose apresentado, foi verificado dentro do conteúdo da análise da equipe médica escolar e pelo Conselho de Medidas de Prevenção da Tuberculose nas Escolas em Toyohashi. De acordo com isto, foi constatada a necessidade de fazer um exame minucioso como segue abaixo. Contando com a compreensão dos senhores, pedimos para que preencham o "questionário para o exame de reação à tuberculina" que segue à parte e entreguem para a escola até o dia 日 ___mês 月 ___ano 年___ . Qualquer dúvida ou inconveniência quanto ao exame indicado, comuniquem à escola.

1 検査内容 Conteúdo do exame

ツベルクリン反応検査 Exame de reação à tuberculina.

胸部レントゲン直接撮影（ツベルクリン反応陽性者のみ実施） Raio X do tórax(somente os alunos que tiveram o resultado positivo no exame de reação à tuberculina).

2 実施日 Data de realização do exame ano mês dia

ツベルクリン接種 Vacinação da tuberculina 平成 ____年 ____月 ____日

ツベルクリン反応判定 Resultado do exame de tuberculina 平成 ____年 ____月 ____日

胸部レントゲン直接撮影 Raio X do tórax 平成 ____年 ____月 ____日

実験結果：検索成功率

3要因(2×2×2)の分散分析

	1分未満	2分未満	3分未満
領域辞書 (参加者内要因)	*** 辞書なし<あり 0.30 0.46	** 辞書なし<あり 0.55 0.75	* 辞書なし<あり 0.74 0.88
サジェスト機能 (参加者間要因)	*** 機能なし<あり 0.29 0.47	** 機能なし<あり 0.56 0.73	---
ユーザ種別 (参加者間要因)	**** 学生<実ユーザ 0.23 0.53	* 学生<実ユーザ 0.57 0.72	---

* p<0.05, ** p<0.01, *** p<0.005, **** p<0.001

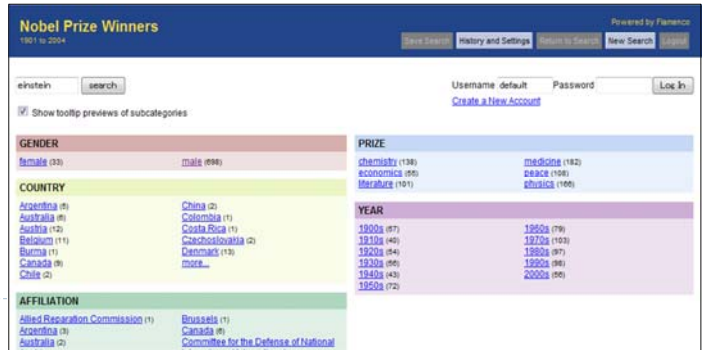
- 1,2,3分未満全ての場合で，交互作用なし (p > 0.05)

ユーザ種別を考慮した評価結果の考察

- 領域辞書およびサジェスト機能の利用はユーザ種別を問わず検索成功率の向上に貢献
- 背景知識を持つ実ユーザの方が検索成功率が高い
  - サジェスト機能によりキーワードの吟味を助けると考えられる
  - 背景知識を持たない場合(学生)はサジェスト機能が検索時間の長時間化抑制に貢献

サジェスト機能とファセット検索の併用

- ファセット検索を統制しない状況での検証
  - ファセット検索とサジェスト機能を組み合わせた場合の効果については未検証
- ファセット検索システムFlamenco
  - ノーベル賞受賞者検索システム



ノーベル賞受賞者のファセット検索サイトを用いた評価実験

- 目的
  - サジェスト機能と領域辞書がファセット検索併用時に検索成功率向上に及ぼす影響の検証
    - 背景知識を与えられた状況での検証
- 実験協力者
  - 情報系学部生32名 (うち女性：13名)
- 課題
  - ノーベル賞受賞者に関するページを検索する課題
  - 正解受賞者の生い立ち，受賞経緯について書かれている文書に空欄を設け提示
    - 前提知識の代替となるような専門用語を文書中に付与
  - 全部で4課題
    - 制限時間は1課題5分間

サジェスト機能とファセット検索を併用した場合の評価結果：検索成功率

2要因(2×2)の分散分析

	1分未満	2分未満	3分未満
領域辞書 (参加者内要因)	* 辞書なし<あり 0.28 0.45	* 辞書なし<あり 0.56 0.75	---
サジェスト機能 (参加者間要因)	* 機能なし<あり 0.25 0.48	---	* 機能なし<あり 0.72 0.91

* $p < 0.05$

- 1,2,3分未満全ての場合で，交互作用なし ($p > 0.05$)

考察： サジェスト機能とファセット検索を併用した場合

- 2分未満で領域辞書利用による検索成功率が上昇
  - 校務文書ポータルサイト実験の結果に概ね一致
- 1分以内及び3分以内ではサジェスト機能の利用による検索成功率が上昇
  - 背景知識を利用できる状況：サジェスト機能によるキーワード候補の吟味により1分以内の短時間で検索が可能
  - 探索が困難な状況：サジェスト機能によってキーワードの想起を助けることで検索を3分程度に抑えることが可能

まとめと考察

- ユーザ種別
  - 背景知識のあるユーザ：領域辞書とサジェスト機能を用いることで，キーワードの吟味に貢献
  - 背景知識のないユーザ：領域辞書とサジェスト機能を用いることで，検索の長時間化抑制に貢献
- サジェスト機能のファセット検索との併用
  - 領域辞書とサジェスト機能はファセット検索併用時も検索成功率の向上に貢献
  - ファセット検索はand検索に代わり，確実な絞り込みの追加を提供

参考文献

- [澤 2010] 澤 晃平，岡野 友輔，堀 雅洋，喜多 千草: 多言語校務文書ポータルサイトのための領域知識を用いたキーワード検索支援.
- [堀 2010] 堀 雅洋，大西奈緒，喜多千草: 多言語校務文書共有のためのポータルサイト構築：カードソートを用いた分類体系の設計と評価．情報処理学会論文誌，Vo. 52, No. 2 (2010).
- [White 2007]White, R. W. and Marchionini, G.: Examining the effectiveness of real-time query expansion. *Information Processing and Management*, Vol. 43, No. 3, pp. 685–704 (2007)
- [Sihvonen 2004] Sihvonen, A. and Vakkari, P.: Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation*, Vol. 60, No. 6, pp. 673–690 (2004)