数式検索システムの実用化に向けて

- 数式検索システムの基本アルゴリズム
- 論文検索などへの応用の検討

RIMS研究集会「紀要の電子化と周辺の話題」

2008年9月3日

大阪大学基礎工学研究科　橋本 英樹，土方 嘉徳，　西田 正吾

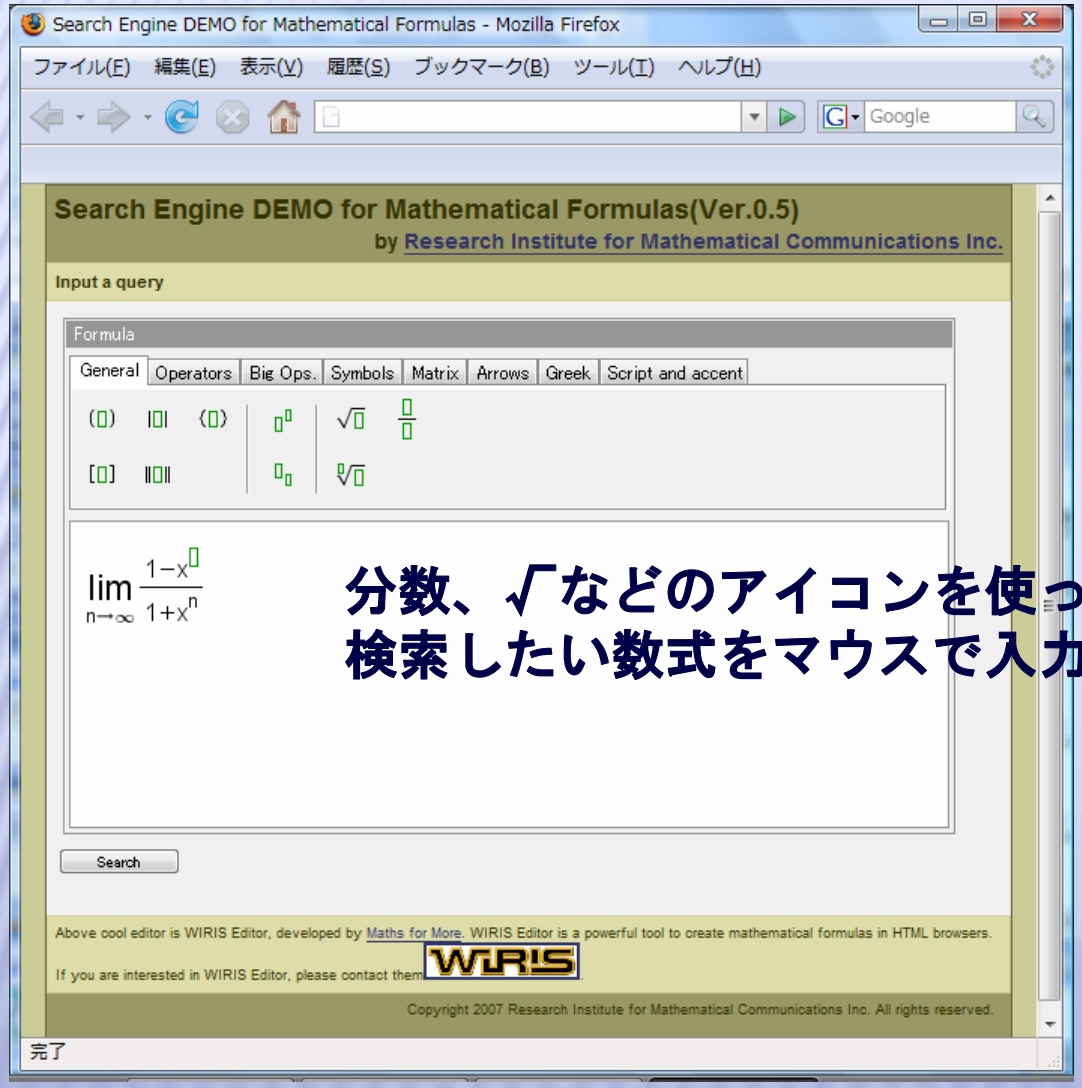
数式に特化した検索

これまでの検索エンジンでは $\int \frac{1}{\sqrt{1-x^2}} dx$ のような数式をキーワード（クエリ）として数式を含む資料を検索することは困難でした。

数式は科学技術や社会現象などの知識を表現する最良の方法です。

数式検索技術を用いた教材やサービスの提供により、学生、教師、研究者、エンジニアを支援し、知的生産性の向上に寄与することを目標とします。

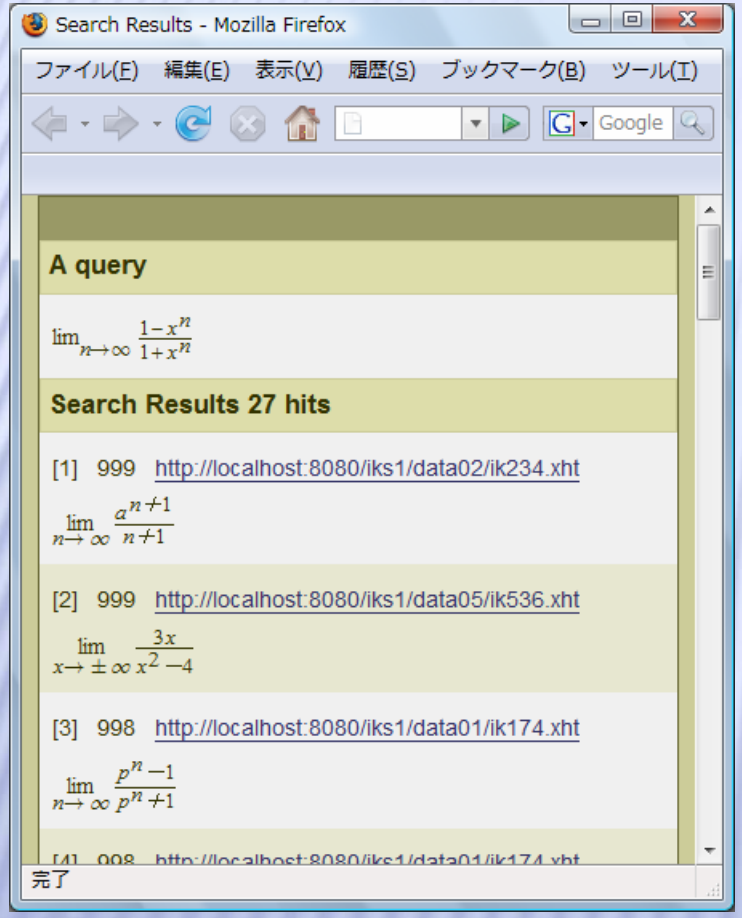
デモンストレーション



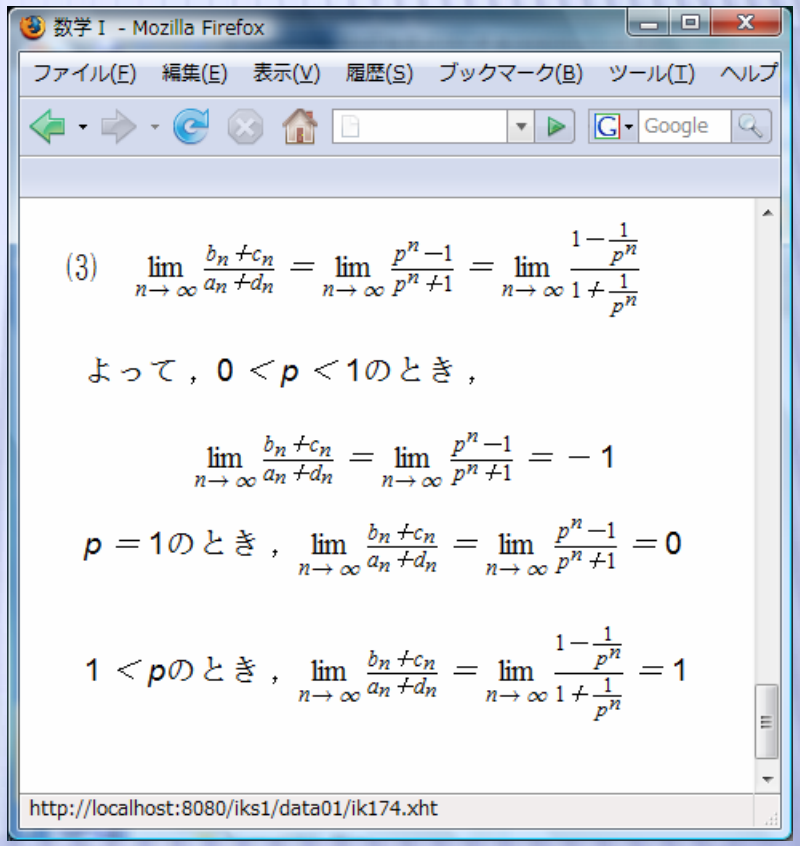
分数、√などのアイコンを使って、
検索したい数式をマウスで入力

デモンストレーション

検索結果表示



見つかった文書を表示



MathMLの利用

- MathMLで記述された数式を含むWebページをブラウザで表示可能
- 例えばインターネットで見つけたMathMLオブジェクトをMathematicaやMapleなどの数式処理ソフトに入力して計算を行うなど、数式データの再利用が容易
- TeX、WordからMathMLへの変換も可能
- OpenDocumentFormat（ODF）でもMathMLを採用

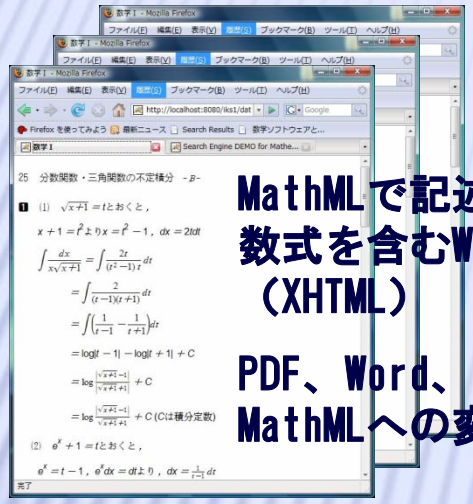
  平成19年3月に各府省情報化統括責任者（CIO）連絡会議において決定された「情報システムに係る政府調達の基本指針」では、ODFファイルフォーマットのようなオープンな国際標準規格の採用を優先するとされています。

- 日本数学会の欧文誌 Journal of Mathematical Society of Japan ではアブストラクトをMathMLで記述
- DLMFプロジェクト（NIST）、Project Euclid、など

- PDF → MathML （InftyProject）

システム構成例

検索対象文書集合



MathMLで記述された
数式を含むWebページ
（XHTML）

PDF、Word、TeXから
MathMLへの変換も可能

検索インデックス
作成プログラム

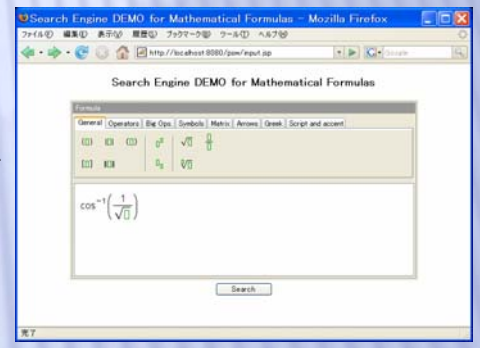
検索サーバー

検索
プログラム

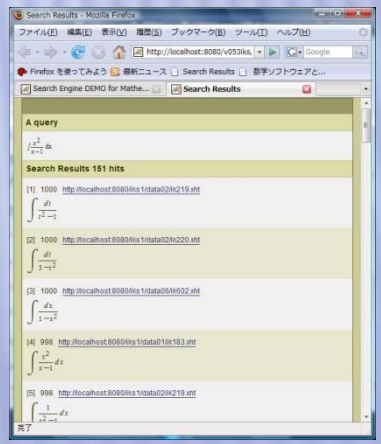
検索
インデックス

ユーザーPC

クエリ入力プログラム



検索結果表示プログラム



（全文）検索エンジンの仕組み

本文
PDF

（文書ID=20）
『東京都の輸出産業の、、、』

『東京』『輸出』で検索

Nグラムを用いた転置インデックス

Nグラム（N=2）	文書ID
東京	1, 18, 20, 49, , ,
京都	20, 103, , ,
都の	5, 20, 21, 75, , ,
輸出	5, 18, 20, 146,
出産	20, 390,
産業	1, 5, 18, 20, 36, , ,

文書ID:
18, 20, ・・・

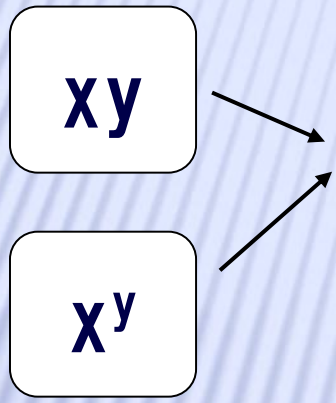
検索結果表示

適合文書
リスト

・この例では『京都』『出産』を検索したとき、、、
・プラットホーム ⇔ プラットフォーム
・パソコン ⇔ PC

数式の構造を反映した検索

日本語の検索では文書中の文字や単語の出現情報を用いて検索を実行する。これに対して、数式の検索では、、、



どちらの数式もｘとｙの２つの記号を含んでいるが、数式の意味と構造は異なる

従来のシステムでは、xとyという記号の出現情報だけを用いて検索を行うため、意図した情報を検索することは困難

数式の構造を反映した検索の仕組みが必要

MathMLによる数式表記

$$\sqrt{x^2 + y^2}$$

MathMLはWeb上で数式を表記するため、
World Wide Web Consortium （W3C)により
勧告された国際規格です

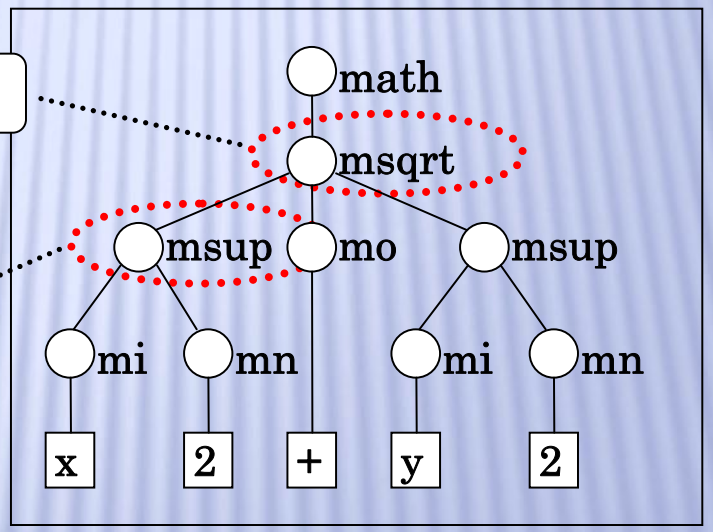
MathMLによる表記

```
<math>
  <msqrt>
    <msup>
      <mi>x</mi>
      <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
      <mi>y</mi>
      <mn>2</mn>
    </msup>
  </msqrt>
</math>
```

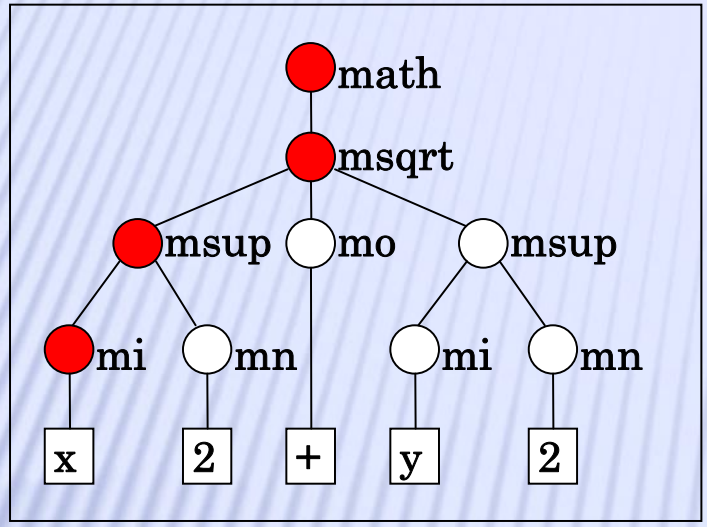
平方根

上付き

DOMツリー



数式の階層構造を用いたインデックス



MathMLオブジェクトのDOMツリーの
ルートから葉ノードまでのパスの
XPath表記をキーとして転置ファイル
を作成

XPath表記	文書IDリスト
/math/msqrt/msup/mi[text()=“x”]	1,52,70,271,,,
/math/mfrac/mo[text()=” ∂ ”]	2,16,55,102,,,
/math/mfrac/mfrac/mi[text()=” π ”]	22,93,181,,,

類似数式検索

検索キー

$$\pi = 2\int_0^\infty \frac{\sin^2(t)}{t^2}dt$$

ヒットした公式

$$\pi = 2\int_0^\infty \frac{\sin^2(t)}{t^2}dt$$

$$\pi = \frac{8}{3}\int_0^\infty \frac{\sin^3(t)}{t^3}dt$$

$$\pi = 3\int_0^\infty \frac{\sin^4(t)}{t^4}dt$$

$$\pi = \frac{384}{115}\int_0^\infty \frac{\sin^5(t)}{t^5}dt$$

$$\pi = \left(2^n \int_0^\infty \frac{\sin^n(t)}{t^n}dt\right) \Bigg/ \left(n\sum_{k=0}^{\left\lfloor\frac{n-1}{2}\right\rfloor} \frac{(-1)^k (n-2k)^{n-1}}{k!(n-k)!}\right)$$

検索結果のソート

- Googleライクなランキング
- 論文の閲覧回数、引用回数によってソートして表示
- 出版年月日
- 目次順

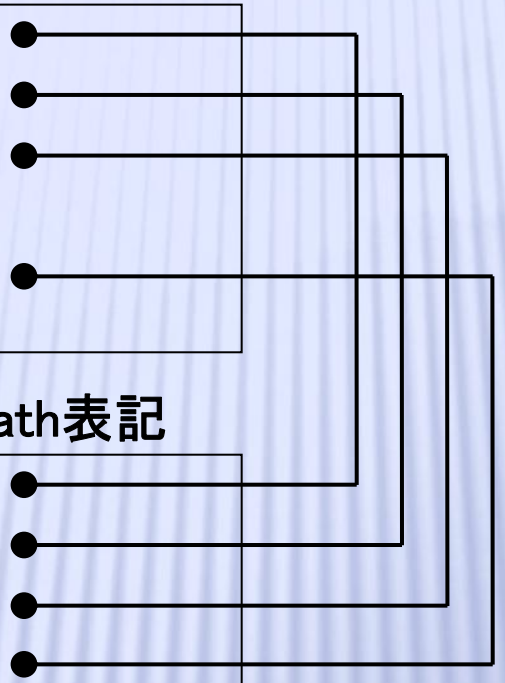
数式の類似度に基づくランキング

x^2+y^2に含まれるパスのXPath表記

math/msup/mi[text()="x"]
math/msup/mn[text()="2"]
math/mo[text()="+"]
math/msup/mi[text()="y"]
math/msup/mn[text()="2"]

$x^2+2ax+a^2$に含まれるパスのXPath表記

math/msup/mi[text()="x"]
math/msup/mn[text()="2"]
math/mo[text()="+"]
math/msup/mn[text()="2"]
math/msup/mi[text()="a"]
math/msup/mi[text()=“x"]
math/mo[text()="+"]
math/msup/mi[text()="a"]
math/msup/mn[text()="2"]



一致するパスが
多いほど
類似度が高い

数式特有の問題への対応

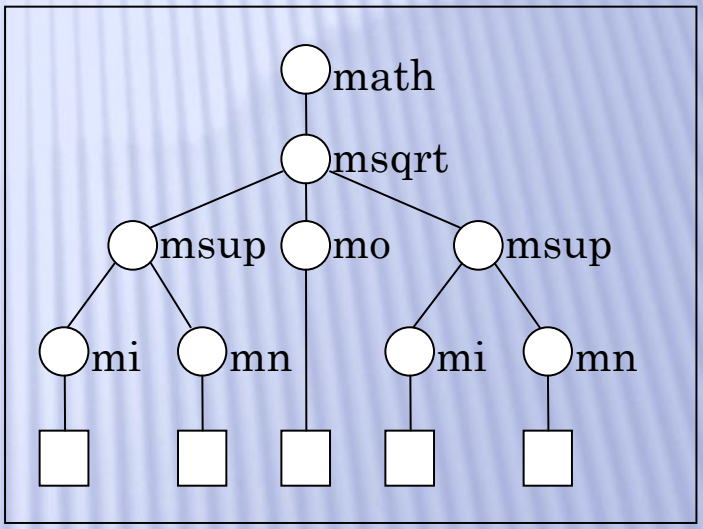
$\sqrt{x^2 + y^2}$

```
<math>
  <msqrt>
    <msup>
      <mi>x</mi>
      <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
      <mi>y</mi>
      <mn>2</mn>
    </msup>
  </msqrt>
</math>
```

$\sqrt{a^2 + b^2}$

```
<math>
  <msqrt>
    <msup>
      <mi>a</mi>
      <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
      <mi>b</mi>
      <mn>2</mn>
    </msup>
  </msqrt>
</math>
```

ツリー構造は同一



ツリー構造だけに着目することにより、
異なる文字で表記された数式の検索が可能

複数の表記法

$$\nabla^2 u \quad \Leftrightarrow \quad \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) u \quad \Leftrightarrow \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

$$\frac{1}{x} \quad \Leftrightarrow \quad x^{-1}$$

複数の表記を同じものとして扱うためのルールを用意して
システムに組み込む必要がある

応用事例の構築

数式検索エンジンだけあっても意味がない。
検索対象となる有用なコンテンツと検索エンジンが組み合わさることによって利用価値が生まれる。

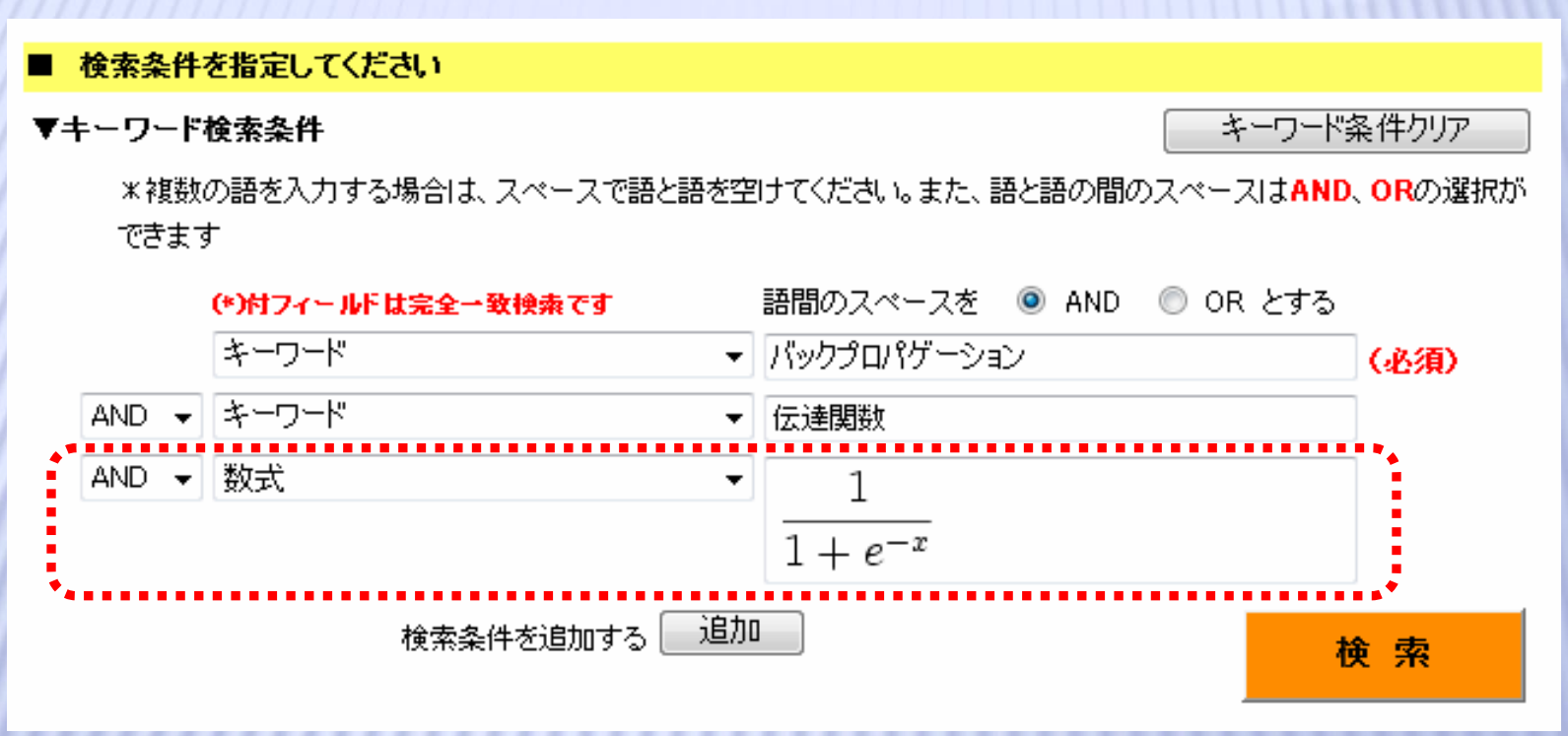
- 論文検索、技術情報検索、特許検索
- 教材検索
- 数学辞典検索、数学公式集検索
- Q&Aサイト

対象分野ごとに
コンテンツの
特徴が異なる
↓
カスタマイズが必要

誰が使うのか？

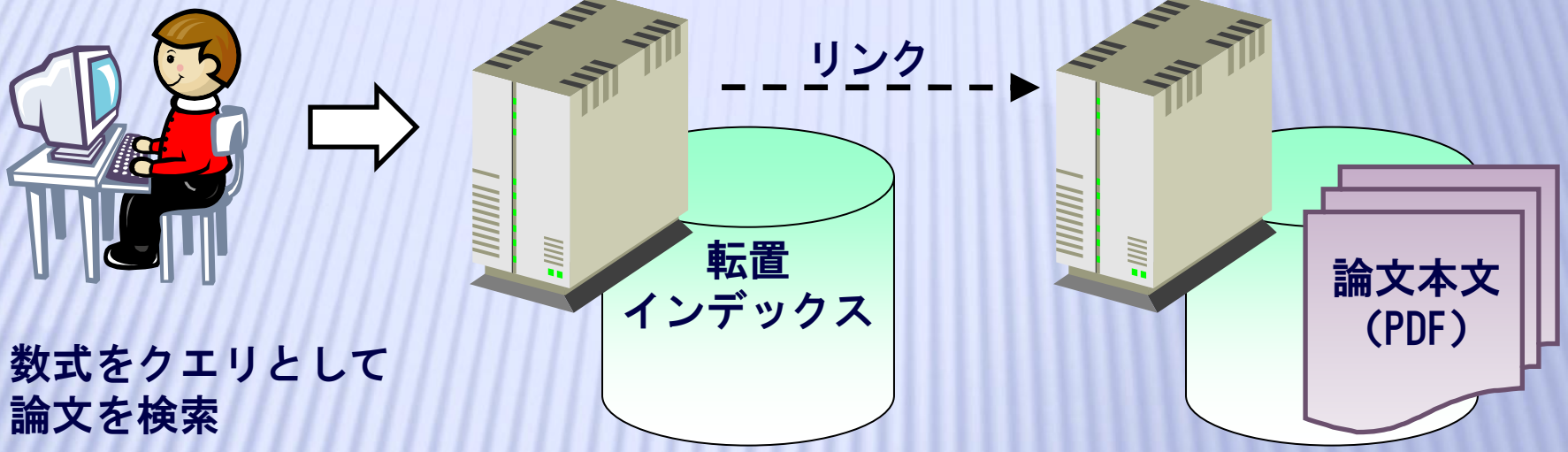
- 分野の専門家、異分野の研究者／技術者
- 学生、教師、教材作成者
- 普通の人が好奇心を満たすために

文献検索サービスでの利用イメージ（将来像）



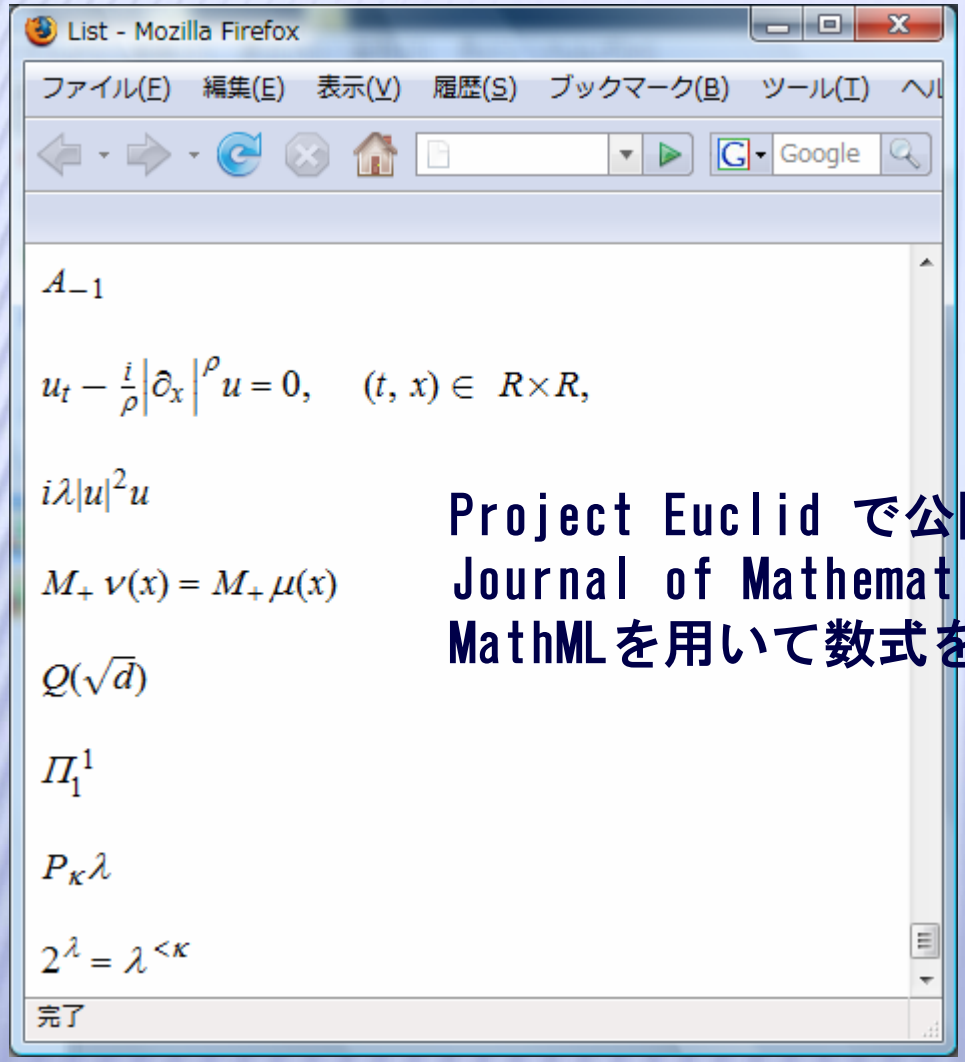
- キーワードと数式を組み合わせた検索によって正確な絞り込みが可能

システム構成例



※論文検索システムは論文公開システムとは別の独立したシステム
（既存の論文公開システムのシステム変更は不要）

JMSJの事例



Project Euclid で公開している日本数学会の欧文誌 Journal of Mathematical Society of Japan では、MathMLを用いて数式を記述したアブストラクトを提供

分野ごとの特色・記号表

流体力学では、		別の分野では、
c	比熱	光速度？
ρ	密度	
τ	応力	時定数？
Re	レイノルズ数	

電磁気学では虚数単位を j で表記

文献サービスに数式検索をプラスすると

- キーワードと数式を組み合わせた検索によって正確な絞り込みが可能

→利便性の向上、省力化、効率化

- 異分野での同一数式の活用事例の発見
- 講師、学生などが学習対象の数式の応用事例を検索

→ユーザー層の拡大

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}=-\frac{v}{\kappa}\frac{\partial T}{\partial x}$$

⇔

この熱伝導のシミュレーションソフトを開発するとき、熱伝導以外の分野で同じような微分方程式の例を調べると良いヒントがあるかも

文献検索サービスに数式検索は必要なのか？

・その分野の専門家以外の人にも利用してもらうために
（キーワードや著者名の知識が少なくても）

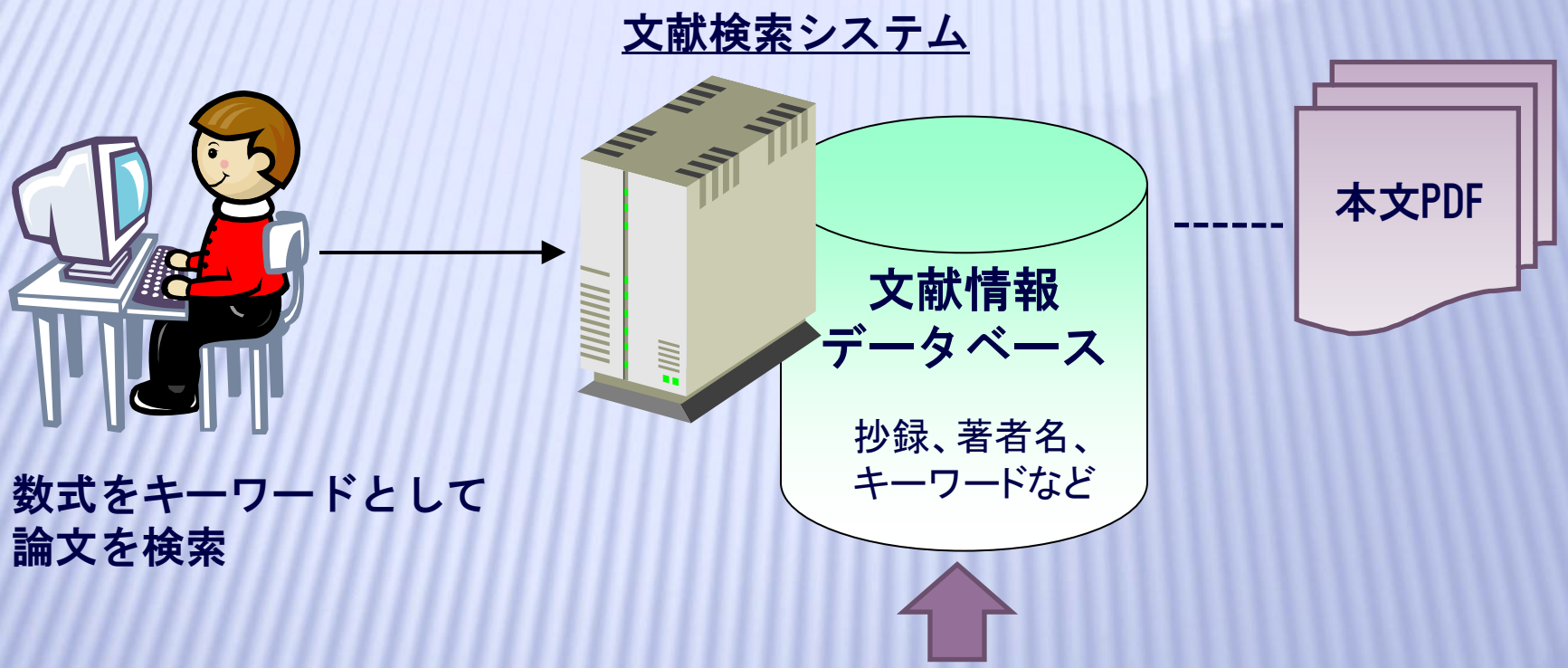
$$\frac{1}{1+e^{-x}}$$ ⇔ 『シグモイド関数』

・より多くの人に論文を見つけてもらい、
知識を活用してもらうために
（社会への還元）

→学術コンテンツ活用、技術情報利用の入り口を広げる

数式をメタデータとして利用

文献情報（2次情報）データベースに数式データを付加



発展

- 査読支援に活用

  →査読時に同じような数式を多く含む既存の論文を表示

- 同じような数式を含む文書を集める（クラスタリング）
- 同じような数式を含む文書を関連付ける（リンク生成）