I-Discover

—Linked Data に基づく
電子情報通信学会文献検索システム—

西野文人 Fumihito Nishino 富士通研究所ソフトウェアシステム研究所

1 はじめに

I-Discover[®] (正式名称, IEICE Knowledge Discovery) は電子情報通信学会 (以降, 信学会) の文献検索システムであり, 本稿が皆様のお手元に届くころには, 既に運用が始まっており, 実際に使ってみたという方も多くおられることと思う。しかし, I-Discover は単なる文献検索システムではない。Linked Data^{(1), (2)} という技術を採用することによって, 情報通信の分野における知識のハブとなることを目指したものであり, 今後の学会の在り方にも深く関与するシステムである。本稿では, まず I-Discover の概観を紹介し, 開発の経緯, システム・デザイン, I-Discover の持つ様々な機能を紹介する。そして, 今後のシステムの改版に向けて検討すべきと思われることを紹介する。

2 I-Discover とは

まずは, 図 1 を見て頂きたい。これは, 検索窓に「誤り制御」という語句を入力して検索したときの検索結果画面である。ここで文献の検索結果だけを見れば通常の検索システムと似ているが, I-Discover では, 文献だけでなく, キーワード, 出版物などのカテゴリ別に, 検索窓に入力された語句に関連する情報が提示される。ここの文献の検索結果から文献を選ぶことも可能であるが, 提示されたキーワードや出版物の中から適切と思われるものを選択することもできる。ここで例えば, 「制御信号の誤り制御」というキーワードを選択すると, 図 2 のようなキーワード表示画面に移り, 選択されたキーワードに関連する情報として, そのキーワードが付与された文献や関連キーワード, 更に選択されたキーワードで外部サイトを検索した結果が表示される (外部サイトとしては CiNii^{*1}, Google^{*2},

IEEE Xplore^{*3}, ACM Digital Library^{*4}, DBpedia^{*5} が選択できる)。

図 3 は図 2 から文献を選択したときの文献表示画面である。選択された文献に関する詳細情報が表示されると同時に, 関連するキーワードの情報や外部サイトの検索結果が表示される。

このように I-Discover は文献検索システムではあるが, Google のような Web 検索システムとは設計思想が異なっている。詳細は後述するが, I-Discover は, 各資源に対して, 例えば文献に対してはタイトル, 著者, 日付など, 人物ならば姓名や所属といった各資源に対して, 資源の特徴を示すデータであるメタデータ^{*6}を検索し, 表示できるようにしたシステムなのである。

図 1 「誤り制御」の検索結果画面

- * 3 <http://ieeexplore.ieee.org/>
- * 4 <http://dl.acm.org/>
- * 5 Wikipedia を Linked Open Data 化したもの。
<http://dbpedia.org/>
- * 6 メタデータは記述対象となる情報資源に対して, 決められた属性の属性値を記したものであり, 基本的な目的は情報資源の発見・記述・管理・蓄積である。ちなみに, Web の父と呼ばれる Tim Berners-Lee はメタデータを「ウェブ資源等に関する機械理解可能な情報」と定義している⁽³⁾。

* 1 日本国内の学会・大学等で刊行した学術雑誌に掲載された論文が検索できる。 <http://ci.nii.ac.jp/>

* 2 <http://www.google.com/>

図2 キーワード「制御信号誤り制御」の画面

図3 論文のメタデータ表示画面

3 I-Scover 開発の経緯

信学会では、論文誌論文、研究会技報、全国大会論文、国際会議予稿など多くの出版物を刊行しており、その大半が電子化されオンラインアーカイブ化も進んでいる。しかしながら、各出版物に対して別個のシステムが構築されていて、全出版物を横断的に検索することができないという問題を抱えていた。

一方、世の中ではデータを特定のソフトウェアの機能に限定されることなく、構造化データとして公開し、その情報にアクセスしたときに有用な情報を取り出せるようにすることで利便性を高めようとする Linked Data という仕組みが注目されつつあった。

そこで、学会の所有するコンテンツに関するメタデータの表現方法であるメタデータスキーマを統一し、そのスキーマに沿って各コンテンツのメタデータを整備し、更にこれらを Linked Data 化することでメタデータの

再利用性を高めるとともに、この Linked Data を検索・閲覧できるシステム I-Scover を構築することで、情報通信分野関係者の利便性を高めることとした。Linked Data 化による具体的なメリットについては後述する。

4 I-Scover システム

図4はI-Scoverの全体構成を示した図である。I-Scoverは論文誌DB、技報DB、大会DB、国際会議DBからメタデータのみを取り出して、まず統一メタデータ形式に変換している。この統一メタデータに対してURI (Uniform Resource Identifier) 付与、RDF (Resource Description Framework) 化等を行い、Linked Data として利用できるようにしている。Linked Data 化されたメタデータは外部データとも連携してデータが収集・再構成され、I-Scover ブラウザを通して検索・ナビゲーションができるようにしている。

4.1 統一メタデータ

学会が保有する情報の中から、論文（文献）、文献の著者（人物）、著者の所属（機関）、文献の在りか（出版物）、会議の開催日や開催場所（イベント）、文献に付与されているキーワード（専門用語）という六つのカテゴリに対してメタデータを整備することにした。まずは、それぞれのカテゴリに対してどのような属性があり得るのかを整理し、統一メタデータ仕様とした。その上で、論文誌、研究会技報、全国大会、一部の国際会議予稿に関して、この統一メタデータへの変換を行った。その際、著者の姓名や所属、会員番号などの情報が不完全なものについては、ほかのデータとの突き合わせ等により解決できるものは解決してメタデータ化を行った。更に、文字列ベースではなく、人物・機関単位で文献が検索・分析ができるようにするために、表記ゆれ等を考慮して同一人物・機関・概念は同じIDになるように名寄せすることに努めた。名寄せに関しては、所属、共著者、キーワードなどの情報を複合的に利用すれば精度の向上は図れるが、それでも100%の精度を達成することは無理であり、時間とコストの関係から、名寄せに関しては次版以降への課題となった。

抽出された文献、人物、専門用語等の各エンティティに対してグローバルに一意な識別子としてURIを与えている。そのURIにウェブブラウザで広くサポートされているHTTP (Hypertext Transfer Protocol) でアクセスしたときに有用な情報を取り出せるようにし、また外部の情報にリンクを張った。各エンティティの情報は Linked Data として一般的であるメタデータを主語、述語、目的語の三つ組の集合として表現するRDFという枠組みにした。この述語として使用する語集もなるべ

図4 I-Scoverの構成

く独自語彙を避け、文書などの資源のタイトル、作者、作成日などの書誌的情報をメタデータとしてするための語彙である Dublin Core^{*7}や、人や組織に関する情報を記述する語彙である vCard^{*8}、人とその活動に関する情報を記述する語彙である FOAF (Friend of a friend)^{*9}といったよく知られた語彙セットの中から選択するようにすることでメタデータの再利用性を高めた。

4.2 I-Scoverブラウザ

抽出されたメタデータは Linked Data として保持されており、これをデータとして活用できるが、一般の利用者に対してはこの Linked Data から所望の文献を検索・閲覧することを支援するシステムの提供が必要である。そのためのシステムが I-Scover ブラウザである。I-Scover ブラウザは一般のブラウザから <http://i-scover.ieice.org/> にアクセスすることによって、I-Scover の持つ Linked Data を検索・閲覧できるようにしたものである。この I-Scover ブラウザの特徴の一つは、ダイナミック・パブリッシングである。すなわち、I-Scover では個々の Web ページを HTML 形式で保持しているのではなく、個々のメタデータからダイナミックに Web ページを生成している。

また、I-Scover では、探す対象が明確になっているものを探す検索 (lookup search) だけでなく、散在している苺を摘んでいくように、漠然とした情報ニーズからウェブ上を探りながら少しずつ思考を加えて自分に必要な情報を拾い集めていく苺摘み (berrypicking) 探索 (exploratory search)⁽⁴⁾ を容易に行えるようにしている。例えば、あるテーマで論文を探す場合、とりあえず思いつく一つのキーワードで論文を探す、検索結果からほ

かのキーワードやその論文が発表された国際会議の情報など探索のヒントになるような情報を見つけてそのヒントを活かして探索を続けることができるように、I-Scover ではヒントになりそうな情報を提示するとともに、そのヒントをクリックできるようにしており、探索が容易に行える。

5 I-Scover の利用

5.1 基本検索

I-Scover の検索は通常の文献検索システムと多少異なるところがある。通常の文献検索システムでは、検索窓に入力された文字列を含む文献を探す。そのとき、複数の文字列を指定することで、それらの文字列の両方を含む文献を探す。これに対して I-Scover の基本となる考えは、まず起点となるメタデータ (例えばキーワード) に焦点を移し、そこに関連付けられている文献を探すというものである。最初の起点となるメタデータの候補の中には指定された文字列をタイトル等を含む文献も提示されるので、一見通常の検索と同様な効果があるように見えるのである。しかし、I-Scover の基本検索では、メタデータの属性値単位で検索が行われる。したがって、検索窓に入力された複数の文字列の全てを含むタイトル等のメタデータを見付けて、そのメタデータを持つ文献を探すことはできるが、タイトルと著者名などの複数のメタデータのそれぞれの文字列を指定して文献を探すような AND 検索は、次に述べるファセット検索を使うことになる。

5.2 ファセット検索(詳細検索)

「ファセット (facet)」とは物事の側面や切り口を意味する言葉であり、例えば文献を探すには著者や出版物名、発行日といったファセット (切り口) に、そのファセットの値の条件を指定することで、求めるものを絞り込むのがファセット検索である (I-Scover 画面では「詳細検索」という名前になっている)。I-Scover では、文献、著者、イベントなどの各カテゴリに対するファセット検索を用意しており、各ファセットの値の条件としては部分文字列一致や日付範囲が指定できるようになっているとともに、それらの条件の AND, OR の組合せを指定した検索ができるようになっている。図4上部はファセット検索画面であり、求める対象のカテゴリを指定するとともに、そのカテゴリに応じたファセットの条件が指定できる (図5の下部は検索結果)。

5.3 外部連携

情報通信分野に関する論文は信学会だけにあるわけで

* 7 <http://dublincore.org/>

* 8 <http://tools.ietf.org/html/rfc2426>

* 9 <http://xmlns.com/foaf/spec/>

The screenshot shows the I-Discover search interface. At the top, there's a header with the I-Discover logo and navigation links. Below that, a search bar is visible. The main section is titled '詳細検索' (Detailed Search) and contains several input fields for search criteria: '検索対象' (Search Target) with a dropdown menu, '著者' (Author) with a text input, 'AND' and 'OR' operators, and '発行日' (Publication Date) with a date range selector. A '検索' (Search) button is at the bottom. The footer contains a disclaimer about the system's purpose and usage.

図5 ファセット検索

はない。そこで I-Discover では、信学会の情報を表示すると同時に、Google、CiNii、IEEE Xplore 等の外部サイトの情報も同時に表示することで利便性を上げている。

5.4 管理者向け機能

多くの人が協力的にデータを追加・編集することで使いやすいシステムを構築できる可能性がある。I-Discover では、キーワードに対して、説明を追加したり、関連するキーワードとリンクを張ったり、あるいは本来同一な人が複数のエントリーに分かれてしまったものを結合するなどの作業を画面上で簡単にできるようにしてある。I-Discover ではシステムをプレビュー環境と本番環境の 2 系統にしておき、編集権限を持つ限られた人のみがプレビュー環境でデータの投入・編集を行い、月に 1 度プレビュー環境のデータを本番環境に移動するという運用が行われる予定である。

6 Linked Data としての I-Discover

6.1 Linked Data の効用

Linked Data は情報を整理し、関連する情報をリンクし、グローバルに利用できるようにする技術であり、様々な分野で活用が進んでいる。I-Discover はそのような Linked Data システムの一つであり、以下のようなことが期待される。

まず、文献、著者、キーワード等の各エンティティは、それぞれ永続的な URI を持つ。例えば、図 3 の論文は <http://i-scovier.ieice.org/iscover/resource/ARTICLE-263FCE21-09B6-6C9D-16FB-8D1A439B37A0> (以降、<http://i-scovier.ieice.org/iscover/> の部分は iscover: と置き換えて、iscover:resource/xxx のように略記する) という URI が割り当てられており、(I-Discover が安定的

に運用されるようになれば) この URI は不変である*¹⁰。したがって、ほかの Web ページから上記の URI でこの論文を直接参照できることになる。それによって、I-Discover 内の各エンティティに直接リンクされることが多くなり、そこから検索・探索として I-Discover を使ってくれる可能性が高まる。専門用語や研究者のページが充実し、情報通信技術分野における百科事典(知識インフラ)の役割を担えるようになれば、文献にとどまらず様々なエンティティに対して直接リンクを張ってもらえることが多くなる可能性がある。そのためには、専門用語の解説を加えたり、関連用語やほかのサイトの辞書にリンクする、業界新聞などのニュースとつなげる、Call for Paper や会議開催などの関連するイベントの情報とつなげるなどが必要である。多くのサイトからリンクされるようになると、検索でも上位にランクされ、サイト訪問者が増加し、論文閲覧数が増加することも期待される。

6.2 データとしての活用

図 3 で示した Web ページは、人が見るために I-Discover が動的に作り出したものであり、実際に保有しているデータはそれぞれのエンティティに関するデータのみであり、それらは RDF で記述されている。実は上記の URI に人がブラウザからアクセスしたときと、プログラムがアクセスしたときとは異なるものが戻ってくるようになっているのである。本稿では詳しいメカニズムは省略するが、人が通常のプログラムから iscover:resource/xxx にアクセスされると iscover:page/xxx に転送し、図 3 のようなページを合成して表示する。一方、プログラムから iscover:resource/xxx にアクセスされた場合は iscover:data/xxx に転送し、RDF をシリアルライズし、図 6 に示すような RDF を XML (Extensible Markup Language) で表現した RDF/XML 形式のデータで返す。RDF/XML 形式は広く通用している形式であり、それを扱うオープンソフトも存在する。したがって、これらのデータを加工して、現在とは別の形式に変換することも容易である。例えば、データの関係をネットワーク図で表示したり、論文データを定まった参考文献の形式に整形したりすることが考えられる。

7 将来展望

2013 年 4 月の初版公開時点ではデータの整備、Linked Data としての特徴の利用の部分に関しては不十分な部分がある。次期版以降に向けて、考えるべき点について以下私見を述べる。

* 10 今後の議論によって URI 付与方針を見直す可能性はあるが、早い段階で固定する予定である。


```

<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:sio="http://www.w3.org/2008/05/skos-xl#"
  xmlns:cc="http://purl.org/ontology/cc/" xmlns:card="http://www.w3.org/2006/04/card#" xmlns:ical="http://www.w3.org/2002/12/cal/ical#"
  xmlns:swrc="http://www.ontoware.org/ontology#" xmlns:prism="http://prismstandard.org/namespaces/basic/2.0/"
  xmlns:ndi="http://ndi.go.jp/ndi/terms#" xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:iscover="http://i-
  scover.ieice.org/terms/iscover#"
  >
  <rdf:Description rdf:about="http://i-scover.ieice.org/iscover/resource/ARTICLE-2C95A19C-4026-DAF7-EEFD-72FF37E2968">
    <rdf:type rdf:resource="http://i-scover.ieice.org/terms/iscover#Article"/>
    <dcterms:title rdf:parseType="Literal" xml:lang="en">
      An error control scheme for the control signal onboard the simplified satellite repeater
    </dcterms:title>
    <dcterms:title rdf:parseType="Literal" xml:lang="ja">簡易衛星中継器における制御信号の誤り制御について</dcterms:title>
    <iscover:authorInfo rdf:parseType="Collection">
      <rdf:Description>
        <iscover:authorInfo rdf:parseType="Resource">...</iscover:authorInfo>
        <iscover:authorInfo rdf:parseType="Resource">...</iscover:authorInfo>
        <iscover:authorInfo rdf:parseType="Resource">...</iscover:authorInfo>
        <iscover:authorInfo rdf:parseType="Resource">...</iscover:authorInfo>
      </rdf:Description>
    </iscover:authorInfo>
    <iscover:author>
      <rdf:Description rdf:about="http://i-scover.ieice.org/iscover/resource/AUTHOR-DA511C67-3270-FA84-C4FF-
        6E28EF9A050">
        <foaf:name xml:lang="en">Takaya Yamazato</foaf:name>
        <foaf:name xml:lang="ja">山崎敦也</foaf:name>
      </rdf:Description>
    </iscover:author>
    <iscover:affiliation>
      <rdf:Description rdf:about="http://i-scover.ieice.org/iscover/resource/ORGANIZATION-46538F7B-63C8-155C-E3B7-
        409F43A5B43">
        <foaf:name xml:lang="en">Nagoya University</foaf:name>
        <foaf:name xml:lang="ja">名古屋大学</foaf:name>
      </rdf:Description>
    </iscover:affiliation>
    </iscover:authorInfo>
    </rdf:Description>
    <iscover:authorInfo>
      <rdf:Description>
        <iscover:authorInfo rdf:parseType="Resource">...</iscover:authorInfo>
        <iscover:authorInfo rdf:parseType="Resource">...</iscover:authorInfo>
      </rdf:Description>
    </iscover:authorInfo>
    <iscover:subject>
      <rdf:Description rdf:about="http://i-scover.ieice.org/iscover/resource/KEYWORD-C83DA570-8BD2-89D4-85E9-34E5990C2FB3">
        <skos:prefLabel xml:lang="ja">パケット衛星通信</skos:prefLabel>
      </rdf:Description>
    </iscover:subject>
  </rdf:Description>
</rdf:RDF>

```

図6 RDF/XML形式

7.1 文字列から意味へ

I-Discover では、文献、著者（人物）などを文字列で扱うのではなく概念で扱えるように名寄せを行い、概念間の関係を整備しているが、その整備はまだ不十分である。メタデータの充実、名寄せを推し進めていくことが必要である。

7.2 高度な検索・分析・可視化へ

RDF で記述されたデータを検索・操作するための言語として SPARQL^{*11} がある。これを利用することで、分散して存在する RDF から必要なデータを見つけて、集計したり表示したりすることができる。例えば、あるキーワードを持つ論文の数を年ごとに集計したり、ほかの Linked Data システムと I-Discover とで横断検索をするようなことが可能になる。SPARQL 検索は万人が利用できるようなものではないので初版の I-Discover では SPARQL 検索を提供していないが、今後どのような形で公開するかも含めて議論が必要であろう。

7.3 ソーシャル情報の活用

様々な学会の文献に対して横断検索を可能とするサイトが存在するのに対して、信学会の論文のみを対象として論文の横断検索システムを提供することの意味は何だろうか。それは学会がその技術分野に対する専門家集団であるということである。それはすなわち各論文に対して評価情報を与えたり、特定の観点に沿って論文やキーワードの集合を設定したり、キーワードの説明や関連性を整備できるということである。今後、専門家集団である特徴を活かした取り組みを考えていくことが必要であろう。

* 11 <http://www.asahi-net.or.jp/~ax2s-kmtm/internet/rdf/rdf-sparql-query.html>

7.4 全体のワークフローの見直し

今回の I-Discover では既存のデータからメタデータを作成した。しかし、論文に付与するキーワードが統制されていなかったり、文字列検索をすることが前提としてタイトルに含まれるキーワードはキーワードとして設定されていないなど、I-Discover がメタデータを利用するという観点では様々な問題があった。今後は I-Discover に登録することを前提として、論文投稿の仕組み等を見直すことが必要であろう。また、全国大会や国際会議等のイベントに関しても投稿、プログラム編成、イベント用 Web ページの作成などの一連の作業の中で、メタデータが自然と整備されるようにすると同時に、そのメタデータを Web ページ作成等に積極的に活用することも考えていくと良いであろう。

8

おわりに

本稿では信学会の文献検索システムである I-Discover システムを紹介した。しかし I-Discover は単に文献を検索するシステムというよりは、知識を取り出すための基盤を目指したものであるそのためには電子情報通信分野における知識が有機的に結合されなければならないが、今回のメタデータ整備はその最初の一步である。今後学会は専門家集団として電子情報通信分野における知の構造化を進め、学術分野、産業分野にとどまらずグローバル社会における学会の価値増大と発展を測っていくことが重要であろう。

文献

- (1) 西野文人, “Linked Data ～つながるデータ, 広がるサービス～,” 信学通誌, no.23, pp.240-244, Dec. 2012.
- (2) トム ヒース, クリスチャン バイツァー (著), 武田英明 (監訳), Linked Data - Web をグローバルなデータ空間にする仕組み (翻訳), 近代科学社, 2013.
- (3) T. Berners-Lee, “Metadata Architecture,” <http://www.w3.org/DesignIssues/Metadata.html>
- (4) M.J. Bates, “The design of browsing and berrypicking techniques for the online search interface,” Online Review, vol.13, no.5, pp.407-424, 1989.

西野文人 (正員)

1981 東工大大学院理工学研究科情報科学専攻修士課程修了。同年、株式会社富士通研究所入所。以来、機械翻訳、情報検索などの自然言語処理に関わる研究開発に従事。現在、同研究所ソフトウェアシステム研究所専任研究員。