5 マイクロソフト社独自開発の MSN Search Engine

浅川 秀治（マイクロソフト（株））
shujia@microsoft.com

Erik Selberg（Microsoft Corp.）
selberg@microsoft.com

本稿では，2005 年に新しくサービスが開始された Microsoft 社の独自開発 MSN Search Engine について概要を述べる．まず，アーキテクチャの上位レベルの説明を行い，続いて基本的な設計目標について説明する．次に，ドキュメントのインデックス化とランク付けに使用した技術について述べる．特に，今回 MSN Search Engine をスクラッチから開発したことで学んだことや，明らかになった問題点に，サービス面技術面の両方の観点から焦点を当て，最後に Microsoft 社の考える今後の検索エンジンの方向性について述べる．Web において最も重要なサービスが検索エンジンであることから，本稿が今後の日本における検索エンジンの開発に役立つことを期待している．

概要

Microsoft 社は 2005 年 2 月 1 日，「Underdog」のコードネームを持つ自社開発のインターネット検索エンジンのサービスを開始した．この検索エンジンは開発に 2 年を費やし，導入時点で 1 億 5 千万検索要求（以下，クエリー）／日の処理能力を保有していた．また，大半のクエリーを 50ms 以下の処理時間で処理しており，これまで Microsoft 社に検索サービスを提供していた検索エンジンプロバイダのサービスを上回る検索効率を MSN Search にもたらすことになった．

本稿では，新しい MSN Search Engine についての概要を説明する．最初に MSN Search Engine の設計目標と上位アーキテクチャについて述べ，続く各章でインデックス化技術としてランク付け技術の概要を説明する．スクラッチから開発した MSN Search Engine の開発では，この間に費やした 2 年間で多くのことを学んだ．この経験から学んだことについて重要なことに特に焦点を当てる．なお，新 MSN Search Engine の日本語バージョンは，2005 年の 6 月末に導入されたが，この日本語バージョンについても貴重な体験を得た．本稿では，日本語検索に固有の問題点についても我々の視点で説明を行い，最後に，将来の検索エンジン開発の進むべき方向についても詳しく述べる．

設計目標とアーキテクチャ

Microsoft 社は 2003 年 1 月，MSN Search のプロバイダであった Yahoo! の Inktomi サービスを自社開発の検索エンジンと置き換えることを目標に自社技術で検索エンジンの開発を行うことを決定した．この時点で，MSN Search は全世界 30 以上の市場をカバーし，1 億 5 千万クエリー／日のピーク処理能力を実現していた．また，クエリーに対して 1 秒以下の応答速度を実現しながらある程度満足のいく検索結果を提供していた．

開発に際しては，大学で研究された技術をベースに小さな設備でサービス開始し，やがてグローバルな商用サービスへと成長した Yahoo! や Google など他の検索エンジン開発会社の開発の歴史とは異なり，新しい MSN Search はプロジェクトの発足と同時に Web に対する網羅性の拡大，膨大な要求を処理することが要求されていた．この実現に向け，以下に述べるとおり，検索システムに対していくつかの設計目標が設定された．

品質と性能

検索エンジンにとって最も重要な品質は，適切な検索結果を迅速に返す検索精度，検索速度である．既存の商用検索エンジンの代表格である Google と Yahoo! はこの点で優れており，MSN Search がこれらエンジンに対して十分競争力を持つための第 1 ステップとして，こ

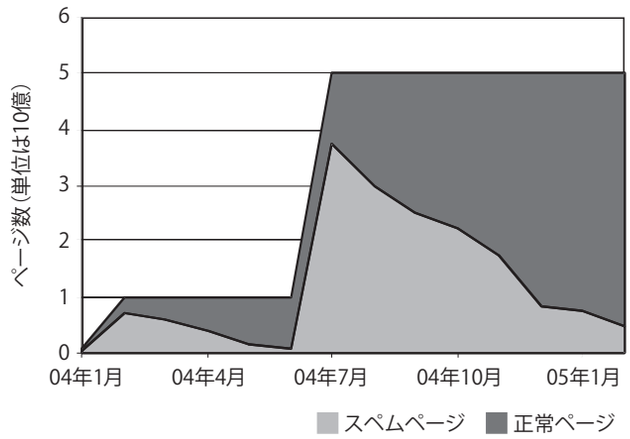

図-1　インデックスサイズとスパム削除ページ量の推移
注）スパムページ数は概略値

れまで Microsoft に検索サービスを提供していた Inktomi 社の検索エンジンをこれら品質面で上回るまで改善を行うことが求められた．適切な検索結果が得られなかったり，検索速度が遅ければ，Microsoft 社のユーザは他のエンジンに切り替えてしまうため当然の目標であった．最優先された設計目標は，検索結果精度の品質を向上させつつ，応答速度を改善することであった．このため，社内において品質基準を設定し，検索結果の精度を改善し，また，検索速度についても厳しい目標を設定して開発を進めた．この時点で，MSN Search には 1 日当たり 1 億 5 千万クエリー，秒当たり最大 4 千クエリーの処理能力が求められていた．これは，秒当たり 4 万にのぼる検索結果が生成されることを意味し，また，検索結果を 250ms 以下，その大半を 50ms 以下で返すことが要求されたのである．

インデックスのサイズと鮮度

2005 年 6 月現在，Microsoft 社の MSN Search Engine では Web 上でクロールした 50 億の最重要ドキュメントに対してインデックス化を行っている．ドキュメントの重要度を測定する方法については，後で述べる．ここで，50 億のドキュメントを選択した 2 つの理由について述べる．

第 1 の理由は，我々が決定を下した時点で Google が 40 億，Yahoo! では 50 億のドキュメントを有しており，このようなライバルのドキュメント保有状況を鑑みるに彼らの保有ドキュメント数で下回ることはビジネス的な競争力の観点から意味をなさないというシンプルな理由があった．第 2 の理由は，高品質な検索結果精度を実現するために，サービス開始日までにスパム対策技術の開発に十分な時間を確保する必要があり盲目的なドキュメントの拡大へのアプローチは避け，品質の改善に優先をおく必要があったことが挙げられる．Web スパムは，その大半がサーバ上で動的に生成されるページであり，検索エンジンによる特定のページのランク付けを不正に水増しすることを目的とするページである．スパムには各種のテクニックが存在するが，その多くが数千から数百万のページを動的に生成することで行われている．現在の検索エンジンでは，スパム対策技術なしでは，インデックスデータベースのサイズに関係なくスパムによって占有されてしまう可能性があり，精度向上を図る上でスパム対策が必須となっている．ドキュメントの保有サイズを増やすために，重要度の低いドキュメントを組み込みながらインデックスを大きくすると，スパムページでないページとスパムページの間の区別が困難になってしまいデータベースの品質が悪化することが考えられる．今回の開発において，インデックスのサイズを拡大すると Web スパムが占める割合も増加し，したがってスパム排除技術を新たに開発しなければならないという発見があった．**図-1** にインデックスサイズを増加した場合におけるスパムページの増加現象を示す．

今後は，スパム対策技術を向上させつつ，Web の成長に合わせてインデックスサイズを拡大していくことを予定している．最近の調査では，Web には少なくとも 110 億のページがあると推定されており [2]，ユーザの検索要求に応えるためにもインデックスを継続して成長させなければならないことは明らかである．

信頼性と堅牢性

これまで述べたような大規模なサービスを提供するためには，数千以上ものサーバが必要になる．このサーバ規模では，PC 製造メーカや社内の内部試験から得られ

た平均故障間隔時間をもとに故障の発生率を計算すると，ディスク故障などによって1日当たり少なくとも1台程度のサーバが失われると予測できた．もちろん，ディスク以外にも，サーバ関連の問題が発生する可能性はある．ユーザは常にサービスが提供されていることを期待しており，マシンの故障がクエリーに影響を及ぼすことがないようシステムを設計する必要があった．Microsoftでは，何らかの理由でシステムの10%が動作不能に陥った場合でも，システム全体として動作を継続できるように設計されている．

オペレーティングシステムとハードウェアの選択

現在，主なWeb検索エンジンのほとんどがUNIXプラットフォーム上に構築されているといっても過言ではない．たとえば，現時点でGoogleなどのエンジンはLinux上で動作しているようである．しかし，Microsoftは，64ビットサーバ向けの最新バージョンを使用するWindowsプラットフォーム上で動作する検索エンジンの開発を選択した．ここで，64ビットWindowsの選択は，Microsoftということで強制されたものではなく，我々にとって最良の選択肢であると考えられ選択された．特に強調しなければならないのは，Windowsプラットフォームが非同期でバッファを使用しない場合のI/Oの最適化においてLinux等のUNIXベースシステムより大きなアドバンテージをもたらすことがある．また，アプリケーションの作成，デバッグ向けのAPIやツールのセットなどを見ても，Linux向けに利用可能な開発ツールと比較するとはるかに多彩なものが提供されており有利であると判断された．

ただし，オペレーティングシステム上のアドバンテージは，その上で開発されるソフトウェアと比較すれば比較的小さなものである．Windowsプラットフォームがもたらす最大のアドバンテージは，開発チームに参加している才能豊かな多数の開発者たちそのものであったと言える．Windows開発者の人材プール化はきわめて重要であるため，チームを短時間で構築することが，2年以下の開発期間でMSN Searchの最初のバージョンを出荷するための条件になったことも大きな要因となっている．

インデックス付けとランク付け

検索の核となる2つの機能は，インデックス化とそこでのランク付けである．これらをどのように処理するかの上位レベルの説明をここでは行う．

インデックスのパーティショニング

設計目標の実現に必要だったのは，インデックスを複数のサーバ間で分担するインデックス化の方式検討であった．Webを対象とした検索エンジンの開発では，時間の経過とともにインデックスサイズが拡大され，また同時に検索するユーザ数が増加するためスケーラビリティの確保は当然必要であり，インデックスはサイズ，クエリー処理量のいずれにおいても容易に拡張可能なものとする必要があった．今回，採用した方式は，現行の検索エンジン業界で標準とも言える方式である．この方式では，ドキュメントごとにインデックスがパーティショニングされ，特定のサーバが全ドキュメントのサブセットの全インデックスを保持する．今回，全インデックスを数千台のサーバにパーティショニングしているが，サーバの選択は任意に行われ，1,000台のサーバが同じ機能を果たすことができるようになっている．クエリーは500台の異なったサーバに送付され，それぞれの結果がマージされて1つのランク付けドキュメントのリストが提供されている．

静的ランクとインデックスの選択

先に述べたとおり最近の調査では，Webサイトから動的に生成されるほぼ無限ともいえるページ数を除いても，110億もの膨大な数のドキュメントがWeb上に存在すると推定されている．このように大規模なWebから50億のページを選択するには，どのような方法があるのだろうか．我々が採用した方法は，クローラが取得したページごとに静的ランクと呼ばれる値を計算しそれに応じて重要度を決定していくというものであった．静的ランクは，全Webページにわたるトータルな順序付けで，今回我々はその上位50億を選択したことになる．本稿ではこれ以上の詳細な説明は行わないが，静的ランクは特定のページへのリンク数やページのコンテンツの状態など，複数の要素を使用して算出されている．

動的ランク

静的ランクがクエリーとは独立したWebページのトータルな順序付けであるのに対し，動的ランクは全体における順序付けがクエリーに依存しているランキング方式となっている．特定のページの動的ランクは，クエリーが特定のページにどの程度強く一致しているかに対応している．動的ランク付けアルゴリズムの例は，現時点でTREC[5]における最強アルゴリズムの1つといわれるbm25[4]である．しかし，Web検索においては，TRECで使用されるのとは大幅に異なるデータが存在するため，

我々が使用するアルゴリズムも bm25 から大幅に異なっている．

Microsoft が現在使用しているのは，RankNet と呼ばれるニューラルネットランキングである．これについては，Burges その他が最近の論文で詳しく述べている[1]．RankNet は MSN Search Engine における特長の 1 つで，優れた検索結果を生成可能であることが証明されている．RankNet では，適切な発見的手法を基本とする複数のコアアルゴリズムが斬新な方法で組み合わされ，より優れた総合的ランク付けアルゴリズムが生成される．我々が使用している適切な発見的手法を基本とするコアの一例は，bm25f である．これは，Microsoft Research Cambridge において Stephen Robertson とそのチームが開発した定番ともいえる bm25 アルゴリズムのバリエーションである．

基本的には，RankNet は，次の関数を使用している．

$$rank = g3\left(\sum_j w32_j\, g2\left(\sum_k w21_{jk}\ x_k + b2_j\right) + b3_i\right) \quad (1)$$

この式において，g2 と g3 は変形関数，w32 と w21 は層 1 から 2，および 2 から 3 への重み付けを表している．また，b2 と b3 は定数である．この関数についての詳細な説明は Burges その他による論文を参照してもらいたい．

ユーザがクエリーを発行すると，クエリーは各インデックスサーバに送信され，そのサーバ上で最良であると考えられる結果のリストが生成される．次に，各サーバからのリストがマージ，ソートされ，総合的なリストが生成される．このリストの上位部分が検索結果として最終的にユーザに提示されるのだが，上位レベルでは静的ランクの組合せからページの動的ランクが計算されている．

総合ランク

特定のページの総合ランクは，静的ランクと動的ランクの組合せにより計算される．今回の検索エンジンでは，広範囲なクエリーや曖昧なクエリーに対しては動的ランクより静的ランクが重要視され，より具体的なクエリーに対しては動的ランクを重視している．たとえば，"Hyatt" で検索した場合は，数千以上に及ぶページに一致し，それぞれのページ間での重要性をクエリーだけから判断することは難しい．したがって，クエリーにかかわりなくページそのものの重要度を活用するために，クエリー "Hyatt" に一致する高い静的ランクを持つページを重視して検索結果とする．おそらく，この場合，米国では www.hyatt.com，日本では www.hyatt.co.jp が上位に表示されることになる．しかし，"Century Hyatt 新宿 " のようなクエリーに対しては，はるかに多くの情報が保持されており状況は異なる．確かに www.hyatt.co.jp が一致するが，新宿にある Century Hyatt の Web ページ（www.centuryhyatt.co.jp）が，たとえ高い静的ランクを持っていない場合でも，ユーザに返される最初の結果でなければならなくなるのである．

関連性の測定

検索エンジンの開発において最も重要な問題の 1 つに，検索エンジンの精度測定がある．今回，明らかになったのは，検索結果の品質を測定するには単独の測定方法だけでは不十分であることが挙げられる．したがって，補完的に複数の測定を使用する必要があった．我々が使用した測定基準の中で優れているのが，Normalized Discounted Cumulative Gain[3] である．検索結果ページにおける 10 番目の位置にあるドキュメント i の NDCG は，次の式で計算される．

$$NDCG_i = N_i \sum_{j=1}^{10} (2^{r(j)} - 1) / \log(1 + j) \quad (2)$$

$r(j)$ は，j 番目のドキュメントの格付け，N_i は選択された正規化定数（完全な順序付けの場合は 1 のスコアが得られる）である．試験やトレーニングの目的で，$r(j)$ は特定の市場（米国，日本など）に適したドキュメントのランキング用に訓練された審査スタッフによりマニュアルで生成されている．

学んだこと

新しい MSN Search の開発，テスト，ベータテスト，サービス開始において我々は多くのことを経験，学習した．既存の検索エンジンサービス提供会社にとっては別に驚くことではないだろうが，今後，何らかの大規模な Web サービスを開発する人の参考にしてもらうためにもこの経験について述べたい．

削除が困難なスパム

スパムページの対応は，高品質の検索を提供するうえで，最も重要であり，同時に，簡単な排除が困難な問題の 1 つでもある．多くのスパマーが，検索結果ページ上での露出機会を増やすため，結果として検索エンジンを経由して 1 人でも多くのエンドユーザを獲得するた

めにスパムページを作成し続けている．スパマーが狙うのは，不法に静的ランクと動的ランクを高いランクとして見せかけることである．近年ほとんど無効になっているテクニックではあるが，静的ランクのスパムに使用されていた簡単なものに動的ドメイン作成がある．たとえば，<sorcier glouton spam example> は特定のドメインに存在するこの種の一例である．現在ではほぼ無効になっているもう 1 つのスパムに，キーワードスタッフィングというものもある．このテクニックは，動的ランクの水増しに使用されている．

このスパム問題への対処を開始したときには，その問題の大きさを改めて実感した．今後も検索結果からスパムを削除するための努力は継続していくが，検索エンジンビジネスの注目度がますます高くなることを考えると，この問題が今後も大きな課題であり続けることは間違いないだろう．

ユーザが望むのは数百万のリンクではなく答

ユーザが検索を実行して検索結果を求めるときに似たモデルに，ATM から現金を引き出す場合が当てはまる．ユーザは，可能な限り短時間で ATM を探し，現金を引き出そうとする．このため必要な行動を取りながら，あちらこちらに動きまわる．検索はこの行為に似ている．でたらめに検索することはなく，何らかの目的のために行動を起こし情報を入手する．しかも，情報をできる限り短時間で入手することが求められる．そうして，情報を入手した後は，その情報を活用して次の行動に移るのである．

求めるドキュメントに到達するために，検索結果として多数のハイパーリンクが提供されることは，ユーザの望むことではない．検索への回答として提供されるドキュメントへのハイパーリンクは確かに役立つものの，ハイパーリンクに変わり回答そのものが提供されるほうがはるかに便利であることは当然である．ただし，ユーザに他の関連リンクを見せたり，関連するクエリーで新たな検索を行うことを働きかけることは，ユーザに別のサービスを使用させるためには効率的な方法であり必要性をすべて否定はできない．

相対的な関連性

Microsoft 社が提供する MSN は全世界 30 カ国以上でサービスを提供しており，これら多数の市場に向けたグローバルサービスを新たに提供するために気付いたことに，検索結果精度がいくつかの個人的な要因によって変化するということがあげられる．特に重要なのは，場所と言語の影響である．たとえば，米国と英国はいずれも英語を共有しているが，多くの場合で，それぞれの国のユーザが希望する検索結果はまったく異なっている．この問題では，英国を，米国と同じように独自の地域として処理し，そのうえで関連性を地域向けに最適化する必要がある．

しかし，地理的に同一地域内でも，検索精度における重要度はきわめて個人的なものと言える．たとえば，クエリー "UW admissions" では，UW は University of Washington，University of Wisconsin，University of Waterloo のいずれかを指しているがどれが正しい結果かは特定できない．さらには，大学内の特定の部門を指す可能性もある．たとえば，コンピュータ学科の学生が University of Washington Computer Science & Engineering の大学院入学ページを探しているかもしれない．ミシガン州の高校生が University of Wisconsin の学部課程入学ページを探している可能性もある．最適のページを決定するためには，ユーザ本人とユーザの検索の目的をよく理解する必要がある．しかし，プライバシーや利便性に関するさまざまな問題がありこれらを簡単に取得することはできない．検索結果が個人ごとに異なるとすれば，特定の情報を見つけ出す方法を，単にクエリーの共有で他のユーザにも活用させることは不可能になる．クエリーを共有する方法は，現在頻繁に行われており，これらの問題を解決するための研究はまだ始まったばかりである．今後の研究においても，この分野が最も重要な項目の 1 つであることは間違いないだろう．

故障を考慮して設計しても困難な作業

MSN Search のアーキテクチャをスクラッチから開発する際に経験したのは，数千の Windows サーバで構成される大規模なシステムを構築するということ，また数多くの理由でマシンが故障することを予期しなければならないということであった．マシンのインストール，障害検出，回復などを自動化するシステムを数多く開発し，またマシンが故障した場合にも機能するコードを開発したのは商用検索エンジンとしては当然のことである．

ディスククラッシュなどの故障への対応は，我々にとって問題ではなかった．しかし，それらに対処するための膨大なエラーやプロセスの処理は，予期した以上に困難な課題であった．たとえば，少数のサーバでマザーボードが原因と思われる断続的なハードウェア障害が発生したことがある．この種の障害ではサーバが完全にクラッシュせず，速度の低下を引き起こしたり，場合によっては誤った結果が生成されてしまう．しかし，修理のためにこれらのサーバをメーカに送っても，修理診断では

故障の原因が特定されず，修理されずにマシンが送り返されることもあった．故障の真の原因がハードウェアにあるのか，測定方法にあるのか，それともソフトウェアにあるのかについて，サーバがオペレーショングループと修理グループ間で行き来することもあった．

日本語に関する問題

MSN Search を日本に導入することは，当初予測していたよりはるかに困難であった．大きな問題がいくつか存在していたが，日本語文字セットをサポートするUTF-8 または同等の符号化を使用して，すべての文字を表現することなどがその一例である．しかし，最も困難な問題は，我々が予測もしていなかったものであった．

語の分割

日本語検索エンジンでは，クエリーをドキュメントや複合語などに一致させるために語を形態素解析（ワードブレーカ）によって分割し，一致する可能性がある語それぞれを識別する必要がある．たとえば，シンプルな例として，米国で人気がある小売店「Bed, Bath, and Beyond」が URL に使用している www.bedbathandbeyond.com がある．検索エンジンが "bed bath beyond" などのクエリーを URL に一致させるには，ワードブレーカが "bedbathandbeyond" を "bed"，"bath"，"and"，および "beyond" に分割しなければならない．適切なドキュメントに正確に一致させるためにはクエリー用語とドキュメントの両方の語を分割する必要がある．しかし，これを適切に実行させることがきわめて困難であることは十分分かっている．細かい分割と粗い分割それぞれにはメリット，デメリットが存在する．たとえば，ドキュメント内の用語 <term-abc> では，細かいワードブレーカは，この語を <term-a>，<term-b>，および <term-c> に分割する．したがって，<term-a> か <term-b> などのクエリーがドキュメントに一致する．しかし，無関係な <term-abd> を含むドキュメントにも一致する可能性が出てしまう．粗いワードブレーカは，<term-abc> を <term-ab> と <term-c> に分割するかもしれない．しかし，この場合，<term-a> や <term-b> などのクエリーはもはやドキュメントと一致しないことになる．

外国の用語

日本では，ひらがな，カタカナ，漢字に加え，数多くの語がローマ字や英語表記のまま使用される．SARS などの頭字語や，企業名では特に一般的である．特に，日本人が英単語を用いて検索を行う場合，日本語のドキュメントが検索結果として返されることが期待される．グローバルサイズの検索エンジンでは，英語クエリーでの検索が，一般的にはより適切であると考えられる他の言語のドキュメントに一致してしまう．たとえば，クエリー "Nikon D70" について考えてみよう．これは，Nikon 製ディジタルカメラ用のクエリーであるが，日本では，Nikon の公式ページは http://www.nikon-image.com/jpn/products/camera/digital/slr/d70/ であり，米国でこれに対応するユーザ向けページは http://www.nikonusa.com/template.php?cat=1&grp=2&productNr=25214 となる．日本人ユーザにはこの英語ページが不要と考えると，この問題に対する簡単な解決法は，日本人ユーザ向けの結果を，JP ドメインか日本語の結果に制限するなど大きなバイアスをかけることである．この方法は一般的な場合には効果があるものの，日本以外にある結果を探している日本人ユーザには大きな問題をもたらしてしまうことになる．たとえば，日本人ユーザが予約の目的でクエリー "Seattle Sheraton" を使用して Seattle Sheraton の Web サイト（www.sheraton.com/seattle）を探しているものとする．単純なバイアスがかけられる場合，"Seattle Sheraton" ではユーザが希望する結果が返されない可能性がある．

日本の Web

MSN Search を含め，一般に Web の検索エンジンは，精度の高い検索を行うために Web の持つ構造を有効使用している．MSN Search でも，Google や Yahoo! と同じように，アンカーテキストを重要視している．ご存知のとおりアンカーテキストは，あるドキュメントから別のドキュメントへのハイパーリンクを記述するテキストである．今回の開発では，日本の Web は構造的に米国やヨーロッパの Web とは異なっている傾向があることを発見した．日本語ドキュメントは他の市場のページほど多くのアンカーテキストリンクを持たない傾向があることが今回の開発の過程で明らかになっている．このため，日本語バージョンの精度向上開発のために最適なページを識別することがより困難になり，代替方法に重点を置く必要が生じたのである．

日本語

日本語を導入するうえで大きな問題の 1 つは技術的なものではなく人的なものでもあった．Microsoft 社の米国本社のある Redmond に本拠地を置く MSN Search の開発チームには，日本語の読み書きができるスタッフが開発当初あまりいなかったのである．たとえば，フラ

ンス語やドイツ語などのヨーロッパ言語の場合は開発者がそれらの言語に堪能でない場合でも，これらの言語には十分な類似性があり，開発チームはこれらの言語に対しては適切な検索結果を生成するうえで確実な進歩をとげることができた．対照的に，開発チームは日本語に苦しみ，この問題を改善するため，Microsoft Japan と密接に協調する必要があり，日米でのチームワークなしには，この製品がサービス提供されることはなかっただろう．最終的には，MSN Search チームにおいて，当然，発生する問題を理解するために，日本語に堪能なスタッフをチームに入れ開発を行った．さらには，開発チームの日本語ができないスタッフは，直面する問題をより理解するために日本語の学習も行ったのである．

この経験から学んだ最大の教訓は，Web 検索に関してはいずれの言語もそれ自体の問題を抱えているということ，そして言語を単に複数の語の集まりとして処理しても適切な結果が得られないということである．

今後の作業

今回の新しい MSN Search Engine は，2003 年初頭から開発が開始され，2 年後の 2005 年 2 月にバージョン 1.0 が米国でサービス開始された．その時点では，リリースの品質としては満足していたものの，希望するレベルにはまだ到達していなかった．サービス開始後，ユーザからの批判の中で最も多かったのは，Google や Yahoo! ほど検索品質が優れていないというものであった．当然，今後も検索エンジンを改良し続け品質向上を図ることはもちろんであるが，我々の目標は単に他の商用検索エンジンである Google や Yahoo! を打ち負かすことではない．我々が努力しているのは，より大きな検索サービス発展という目的のためであり，以下の節で，これらについて述べる．

ユーザの疑問に答える

検索に対するコアミッションは，ユーザの疑問に答えることである．これは，検索エンジン上で最適の検索結果として URL のリンク集をクエリーに対して提供することだけを意味するものではない．大切なのは，ユーザがいくつものリンクにアクセスすることなく，疑問に対する答を直接得られるようにすることであると考えている．つまり，検索結果において，即時に検索の目的とする情報を提供することである．たとえば，1.0 リリースでは，直接的な答があると判断した場合，百科事典データベースである Encarta から得られた検索結果を検索結果上位に示すことで直接ユーザの疑問を解決するようにしている．たとえば，「日本の人口」というクエリーに対しては，検索結果上にこの答をインラインで直接提供している．

広範囲な選択

これまで，Web 検索では Web ドキュメントのインデックスのランキングが重視されてきた．スパムや無用なコンテンツに対するランキングをより小さな値に抑えながら Web 上のすべてのドキュメントに対して適切なランキングを与えるよう努力するのはもちろんであるが，Microsoft 社では他のコンテンツの利用にも期待を寄せている．その一例として，現時点では Web 上で利用できないコンテンツのランキング化がある．これについては，Google や Yahoo! も同じ目標に進んでいると考えており，たとえば，Google は各種ライブラリのプライベートコレクションのインデックス付けでイニシアティブをとっている．

さらには，ユーザが自身で使用するためにユーザ自身のデータにインデックスを付けることにも関心を向けている．Microsoft 社が最近導入した MSN Desktop Search では，ユーザが自身のコンピュータ上のドキュメントや電子メールを効率的に検索することを可能にしている．

シームレスな統合

多くの場合，ユーザは電子メールを書いたり，ドキュメントを作成している際に情報を得るために検索エンジンを活用する．ユーザは Web ブラウザを立ち上げ，検索エンジンに移動し，クエリーを入力し，ドキュメントを調べ，その後でこれまで行っていた作業を続ける．Microsoft の考える目標の 1 つは，ユーザがアプリケーションを変更することなく必要とする情報を得ることができるよう検索を統合することである．つまり，検索をユビキタスなものにし，複数のエントリーポイントから利用できるようにすることである．これを実現するため，Microsoft は検索関連製品をいくつかリリースしている．その 1 つが，Windows Desktop Search であり，ユーザのデスクトップや Outlook に検索へのエントリーポイントを設けることができる（**図 -2**）．このツールを使えば，どのようなアプリケーションを使っているときでも，即，検索を実行できると同時に，ローカル PC 上のファイルの検索から Web の検索というシームレな検索も行える．さらに，ローカルファイルの検索がインデクスを作成して行うことで非常に高速化され，加えて画像や Office ファイルのプレビューを結果画面上で見ることができ（**図 -3**）大幅な検索作業時間の短縮が可能となっている．そうしてもう 1 つ，Desktop Search とともに提

図-2　Desk Top Searchへの検索入力の例

図-3　DeskTop Searchの検索結果画面の例

供しているMSN Toolbarがある．これは，MSN Searchに簡単にアクセスするためのInternet Explorerツールバーで，www.msn.co.jpのような検索サイトに移動することなくクエリーを入力することが可能になるものである．

今後ますますシームレスな検索の必要性が出てくると考えるが，Webからローカル，テキスト，HTMLファイルだけでなく，イメージ，Videoファイルというシームレス化，PC，携帯といったようなシームレス化の拡大も図られる必要がある．さらに大きなレベルでは，Webと現実世界のシームレスなつながりとしてWebと身近かな情報へのシームレス化ということを検索を通して行うこともあるだろう．Microsoft社は今後一層このシームレスな検索体験の提供を模索していく予定である．

プラットフォームとしての検索

現時点でのWeb検索は，検索サービスの発展段階の中では一局面を実現しているに過ぎない．Microsoftが現在考えているのは，斬新で意味のあるアプリケーションがその上で構築できるようなプラットフォームとしての検索サービスである．我々はすでに，その実現に向かって歩み始めているのである．近い将来，それが研究開発の主要分野の1つとなり，検索が新しく斬新な方法でアプリケーションに不可欠なコンポーネントとなることを期待している．この意味で近い将来導入されるものに，各種の検索APIがあげられる．このアプローチでは，検索エンジンは単に情報を得るための目的でアクセスするサイトの1つではもはやなくなり，ユーザ自身のアプリケーション内に組み込まれたWebサービスとして，ユーザが新しいMSN Searchを使用できるようになるはずである．

結論

本稿では，2005年に登場したMSN Search Engineを紹介した．エンジンの設計目標とアーキテクチャに焦点を合わせるとともに，Webページをインデックスし検索結果をランク付けする方法を簡単に説明した．また，このエンジンの開発で学んだこと，特に日本語バージョンのサービス化で学んだことも紹介している．最後に，可能な限り広範囲な情報でユーザの疑問に答える検索エンジンを作成し，またその上に他のアプリケーションが構築できるようなシームレスな統合とプラットフォームを開発するためにこれから検索サービスの開発者が進まなければならない方向について述べた．

参考文献

1) Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G.: Learning to Rank using Gradient Descent. Proc. 22nd ICML.
2) Gulli, A. and Signorini, A.: The Indexable Web is More than 11.5 Billion Pages. Proc. 15th WWW Conference (2005).
3) Jarvelin, K. and Kekalainen, J.: IR Evaluation Methods for Retrieving Highly Relevant Documents. Proc. 23rd ACM SIGIR (2000).
4) Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M. and Payne, A.: Okapi at TREC-4, in NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)(1995).
5) Vorhees, E. M. and Buckland, L. P.: NIST Special Publication 500-261: The 13th Text REtrieval Conference Proceedings (TREC 2004)(2005).

（平成17年7月11日受付）