■7 群（コンピュータソフトウェア）－ 6 編（情報検索とデータマイニング）

1 章 テキスト情報検索の基礎

（執筆者：林　良彦）［2009 年 4 月 受領］

■概要■

Web サーチエンジンが日常的に用いられるようになり，「検索」という用語もすっかり一般的な言葉となった．今では，多くの人々が日々，Web サーチエンジンを用いて「情報検索」を行っている．情報検索（IR：Information Retrieval）とは，情報を蓄積し，後にユーザの要求に応じて対応する部分を効率良く取り出すことをいう．この概念の指す範囲は非常に広いものであるが，

1. どのような情報を蓄積の対象にするか
2. どのようにして情報を蓄積するか
3. どのようなユーザの要求を受け入れるか
4. ユーザの要求に対応する「部分」とはどのようなものか
5. どのようにして対応する部分を効率良く取り出すか
6. 取り出した情報（の部分）をどのようにユーザに提供するか

といった観点から細分類されることになる．

本章で扱うのは，テキスト情報を蓄積・検索の対象とするテキスト情報検索に関する基本的な事項である．Web 検索を対象とする Web サーチエンジンを構成する諸技術は 2 章で詳しく説明される．また，どのようなユーザの要求を受け入れ，どのような情報の単位を扱うか，といった観点からのバリエーションに関するテキスト情報検索の関連技術については，3 章の各節で各論的に解説される．

【本章の構成】

まず 1-1 節では，テキスト情報検索技術の歴史的な流れを観点を絞って振り返り，基本的な用語の定義・説明を行う．また，テキスト情報検索技術の発展に重要な役割を果たした TREC（Text Retrieval Conference）について説明する．次に 1-2 節では，情報検索システムの評価方法について，特に検索の有効性（effectiveness）の評価尺度について説明する．続く 1-3 節では，ユーザにより発せられる検索質問と文書の適合度（relevance）の計算方法，そのために重要となるタームの重み付け（term weighting）などのテキスト情報検索の基本手法について説明する．

ここで，情報検索に関する代表的な教科書のいくつかを紹介しておく．“Introduction to Information Retrieval” [1] は，情報検索に関する最新の教科書で，取り扱っているトピックも包括的である．また，アルゴリズムの記述や必要に応じて数値例も提示されており，本書を一通り読み進めることで情報検索の研究開発のスタートラインに立つことができると思われ，現時点でまず第一にお勧めしたい成書である．“Information Retrieval. Algorithm and Heuristics (Second Edition)” [2] は，情報検索の技術を基本的な技術と周辺・支援的な技術に分けて論じているところが特徴であり，また，多数の数値例が提供されていることが特色であるが，索引が貧弱であるという問題がある．“Modern Information Retrieval” [3] は，検索のユーザインタフェースや情報の視覚化について触れているところ，テキストだけでなくマルチメディア情報の検索にも触れているところに特色がある．日本語で書かれたものでは，“情報検索と言語処理” [4] がある．本

書は，情報検索の基本技術をコンパクトに解説しつつ，情報検索における自然言語処理技術の利用やユーザインタラクションの側面に関して，一定の紙面を割いて述べているところに類書にない特長がある．

■参考文献

1) C.D. Manning, P. Raghavan, and H.Shütze：Introduction to Information Retrieval, Cambridge University Press, 2008.
2) D.A. Grossman and O. Frieder：Information Retrieval, Algorithm and Heuristics (Second Edition), Springer, 2004.
3) R. Baeza-Yates and B. Rieiro-Neto：Modern Information Retrieval, Addison Wesley. 1999.
4) 徳永建伸：情報検索と言語処理，東大出版会, 1999.

■7 群－6 編－1 章

1-1 テキスト情報検索技術の歴史的流れ

（執筆者：林　良彦）［2009 年 4 月 受領］

情報検索の概念は，“information retrieval”という用語 *1 が生まれるより少し前の 1945 年に Bush [1] により提案された memex という仮想の機械に示されていると言われている．memex（Memory Extender）とは，各個人が様々なタイプの情報を蓄積することができ，蓄積された情報に柔軟かつ高速にアクセスすることができる機械である．初期の情報検索においては，学術的な文献情報の蓄積・検索が主な対象であったが，Web の普及とともに様々なタイプの情報がその対象となり，また，検索・アクセスの手段の高度化も進んできている．この意味で，技術は memex が提示した概念に近い方向へ進んできているといえよう．

Lesk による情報検索の歴史に関する解説 [2] では，情報検索の歴史を人の一生に例え，1945 年から 2010 年までを 7 つの年代に区切って詳細に論じている．各年代における時代背景や技術的な問題意識についてはこの解説にゆずり，ここではテキスト情報検索を対象に，以下の 3 つの観点に絞って技術の流れを簡単に整理しておく．

1. 実際に検索の対象になるのは何か：文書に関する情報か/文書の内容そのものか
2. 検索のために使われる語彙はどのようなものか：あらかじめ決められた語彙か/自由な語彙か
3. ユーザはどのように検索の要求条件をシステムに与え，システムはどのように検索結果を返すか

1-1-1 一次情報／二次情報

情報検索の初期においては，文献検索の性格が強かった．すなわち，文献の著者，タイトル，ジャンル分類，発行年，発行場所のような，いわゆる文献の目録情報が主な検索の対象であった．言うまでもなく目録は，一次情報（primary information）である文献の内容そのものではなく，文献に関する情報，すなわち，二次情報（secondary information）である．しかしながら，情報検索の対象が新聞記事や近年の Web 上の情報のように変化するにつれ，一次情報そのものが検索の対象とされるようになった．この背景には，記憶装置を中心とする計算機ハードウェアの進展があり，大量の一次情報をそのまま蓄積することが可能となったという要因もある．

ただし，二次情報の重要性が決して損なわれたわけではなく，特に近年では，セマンティック Web の文脈の中で機械が理解できる（machine-understandable）メタデータ（metadata）の重要性が叫ばれている．メタデータとはデータに関するデータであり，もちろん二次情報に相当する．

1-1-2 統制された語彙／自由な語彙

上記のように，初期の情報検索では文献目録のような二次情報が検索の対象であり，文献の内容に関する情報は，文献のジャンル・トピックを表すものとして付与された索引語（index term）に限られた．また，文献の内容を分析して索引語を自動付与することも困難であったため，統制語彙（controlled vocabulary）を用いて人手によってこれらの索引語を付与することが

*1 この用語が初めて使われたのは 1950 年のことと言われている．

行われた．一方で，シソーラス（thesaurus）を利用することにより，「概念的により広い範囲/狭い範囲」を表す索引語への展開などが試みられた．また，ユーザによる検索質問においては，これらの語を論理的に組み合わせた検索式が利用された．

しかしながら，情報検索が対象とする情報の種別が広がり，情報検索の実際の対象が二次情報から一次情報へと移るにつれ，文書に含まれるタームをそのまま用いること，また，ユーザも自由な語彙を用いた検索質問を発することが主流となった．このような方法論は，それまでの統制語彙とシソーラスの組み合わせによる検索よりも多くの場合で有効とされ，今日の Web 検索においても主流な方法論となっている．このような背景には，自由なテキスト（free text）を処理するための言語処理技術の進展も関与している．

さてここで，いくつかの概念・用語の説明を行っておく．まず，文書（document）とは情報検索の基本的単位である．文書は論理的な単位であり，物理的なファイルの単位とは必ずしも一致しなくてよい．また，ターム（term）とは，単語，または，単語と同等に扱うことが望ましい複合語や句を指す．更に，検索質問（query）とは，ユーザが情報検索システムに与えるデータであり，ユーザが知りたいと考えている事柄をタームの集合や自由なテキストとして具体化したものである．

通常，情報検索のユーザは，自らが達成したい目的に向けて，自らが持つ情報・知識が不十分である（ASK：Anomalous State of Knowledge）[3] という認識を動機として情報検索を行う．このようなユーザの状態を情報要求（information need）という．すなわち，検索質問はユーザが漠然として持っている情報要求を具体化したものであると言える．

1-1-3 検索質問と情報検索結果の形式

ユーザが与える検索質問と情報検索システムが返す検索結果の間には深い関連がある．一般にユーザが少ない情報しか与えなければ，システムが返す結果もそれを反映したものにならざるを得ない．ところが，複雑な情報要求を自由なテキストの形で情報要求としてシステムに与えても，それに見合う情報が返されるとは限らない．情報検索における研究開発のある部分は，このような問題意識のもとになされてきたと言える．例えば，質問応答（QA：Question and Answering）と呼ばれる情報検索の一形態では，知りたいことを具体的に記述した検索質問を与え，文書や文献ではなく，知りたい情報そのものを返そうとしている．

さて，情報検索の初期の段階では，統制語彙の中から選ばれた語を AND/OR/NOT の論理記号で結合した検索式を検索質問として受け付け，この論理的制約を満たす文書集合を返す，いわゆるブーリアン検索（Boolean information retrieval）と呼ばれる形態が主流であった．今日でも特許文献の検索システムなどではブーリアン検索が用いられているが，論理的に明確な基準で検索結果が得られるという利点の一方で，必要十分な情報量を得るための検索式の調整が難しいという問題がある．また，より重大な問題として，論理的な検索式は検索結果が満たすべき制約を与えるだけなので，ユーザの情報要求への適合の度合を加味した文書のランク付けができないという問題がある．

上記で述べたように，文書の一次情報を情報検索の実際の対象とする自由な語彙による検索が主体になるにつれ，ユーザにより与えられる検索質問と文書の間の適合の度合い（relevancy）を評価し，これに基づいて検索結果の文書集合をランキングする形態が主流となった．このような，ユーザがアドホックに発する検索質問に対して適合する文書を文書集合から検索するタ

スクを特にアドホック検索（ad-hoc retrieval）と呼ぶ．本章で説明の対象とするのは，テキストメディアによる情報を検索対象とするアドホック検索のために必要な基本的な事項である．

1-1-4 TREC

テキスト情報検索技術の進展，ならびに，その評価手法の洗練化に関して大きな貢献をなしたのが TREC（Text Retrieval Conference）[4] である．TREC は，米国 NIST（National Institute of Standards and Technology）が主催する情報検索に関するワークショップであり，1992 年の初回以来，毎年 1 回開催されている．

TREC の最大の特徴は，コンテスト型のワークショップであることである．すなわち，主催者が準備する文書集合（test collection），検索課題（topics）に対して，各参加者が自システムによる検索を実行し，その結果を主催者側に提出する．主催者は，提出された検索結果を共通の評価指標によって評価する．また，このために必要な正解データを準備する．このように，共通のデータに対し共通の指標で評価を行うことによって，ワークショップの場では，検索手法の比較，それに基づく技術的議論が交わされる．TREC で用いられる評価指標は，広く情報検索技術の標準的な評価尺度として利用されており，この意味でも TREC の果たした役割は非常に大きい．

TREC では，時代とともに様々なタイプの情報検索を対象とする分科会（Track と呼ばれる）が設定されている．例えば，執筆時点において直近の TREC 2009 においては，Web 検索，ブログ検索などの検索対象に応じた分科会が設定されている．これらの中では，Million Query Track と呼ばれる分科会が興味深い．従来の TREC アドホック検索においては高々50 個の検索課題を対象としていたのに対し，これを大幅に増やす Million Query Track と呼ばれる分科会が行われる．

■参考文献

1) V. Bush：“As We May Think,” The Atlantic Monthly, vol.176, no.1, pp.101-108, 1945. http://www.theatlantic.com/doc/194507/bush
2) M. Lesk：“The Seven Ages of Information Retrieval,” International Federation of Library Association and Institutions, UDT Occasional Paper #5, 1995. http://www.ifla.org/VI/5/op/udtop5/udtop5.htm
3) N.J. Belkin：“Anomalous State of Knowledge,” *in* K.E. Fisher, S. Erdelez, and L. McKechinie（Eds.）：Theories of Information Behavior, ASIST Monograph Series, pp.44-48, 2005.
4) E.M. Voorhees and D.K. Harman：TREC, Experiment and Evaluation in Information Retrieval, MIT Press, 2005.

■7 群－6 編－1 章

1-2 情報検索システムの評価方法

（執筆者：林　良彦）［2009 年 4 月 受領］

次節でテキスト情報検索の基本手法を説明する前に，情報検索システムの評価方法についてまとめておく．情報検索システムは，テキスト情報を対象とするテキスト情報検索システムに限らず，以下の 2 つの観点から評価される．

- 効率性（efficiency）
- 有効性（effectiveness）

ここで，効率性とは検索をいかに少ないコストで行えるかに関する．コストには，要する時間やかかる費用などの様々な観点が考えられ，また，これらは検索システムのユーザビリティといった要素にも大いに左右されるため，共通的な評価指標を定めることが難しい．一方，有効性とは検索がいかに有効であったか，言い換えれば，どの程度「有用な」文書を検索できたかに関する．ここで，文書（document）とは検索の対象となる単位を指し，ファイルなどの物理的な単位とは必ずしも一致しない．有効性に関する評価指標は情報検索の歴史における早期の段階から議論され，特に TREC のコンテストを通して，その運用方法も含めて洗練されてきた．以下では，有効性に関する評価指標に関して説明する．

1-2-1 精度と再現率

ユーザの検索質問に対して検索された文書の中で，検索質問にマッチする文書を適合（relevant）文書，マッチしない文書を不適合（non-relevant）文書と呼ぶ．また，ある検索質問に対してある文書に上記の判断を付与することを適合度判定（relevance judgment）の付与という．

情報検索において最も基本的な有効性の評価指標は，精度（precision）P[*2] と再現率（recall）P である．精度とは，システムによる検索結果である文書集合中に適合文書が占める割合である．一方，再現率は，本来検索されるべき文書のどれだけをシステムが検索できたかの割合である．すなわち，前者は検索質問に適合しない文書の検索をどの程度避けることができるかの指標であり，後者は検索質問に適合する文書をいかに漏れなく検索できるかを示す指標である．一般に，この 2 つはトレードオフの関係にある．例えば，検索質問に対して検索対象のすべての文書を検索結果とすれば，再現率は 100 %となるが，適合率は低いものとなる．このような 2 つの値を総合する指標としてよく用いられるのが F 尺度（F-measure）である．F 尺度は，精度と再現率の調和平均により計算されるが，特に両者を同じ重みで平均することが多く，$F = 2PR/(P + R)$ により計算される．

1-2-2 テストコレクション

精度と再現率により情報検索システムの有効性を評価するためには，テスト用のデータセットが必要である．すなわち，検索対象の文書集合，テスト用の検索質問の集合，各検索質問に対する適合文書の集合を準備する必要がある．テスト用のデータセットのことをテストコレクション（test collection）という．テストコレクションの作成においては，検索対象の文書集合

[*2] 適合率と呼ばれることも多い．

や検索質問の集合を準備することももちろん重要であるが，適合文書の集合を適切に作成するかが特に重要である．テストコレクションにおける各検索質問に対する適合文書集合を準備するためには，人手による適合度判定の付与を必要とするが，対象とする文書集合が大きくなるとこの作業は非常に困難なものとなる．このため，プーリング（pooling）と呼ばれる技法が用いられる．プーリングとは，TREC などの情報検索のコンテストにおいて採用されている方法であり，コンテストに参加する複数の検索システムによる検索結果の和集合を求め，これに含まれる文書のみに対して適合度判定を付与する．プーリングは現実的な方法ではあるが，どの検索システムにおいても検索されない文書は，例え本当は適合文書であったとしても適合度判定の付与の対象とはならず，よって不適合文書とみなされることになるという問題をはらんでいる．

1-2-3 精度-再現率曲線

次節で説明するように，現在の情報検索システムは検索質問に対する適合度に応じて検索された文書をランキングする．例えば上位 10 件までといったランキングの上位までにおいては，精度は高いことが期待できるが，再現率は低いとレベルにあると予想される．一方，ランキングの相当下位までを考慮すれば，再現率は高くなる一方，精度は低くなる．そこで，精度を再現率の関数としてとらえ，横軸に再現率をとり，縦軸に精度をとったグラフを描くことにより，情報検索システムの性能評価を行う．このようなグラフを精度−再現率曲線（precision-recall curve）と呼ぶ．実際の情報検索においては再現率を直接コントロールすることはできないので，検索結果の各ランクごとに精度と再現率を求めることになるが，これを直接グラフにプロットすると，階段状の形状が得られる．そこで，次式によるスムージングを行うことによって，再現率があるレベル（recall level）r にあるときの補間精度（interporated precision）$P'(r)$ を求める．

$$P'(\mathrm{r}) = \max_{r' \geq r} P(r) \qquad (2・1)$$

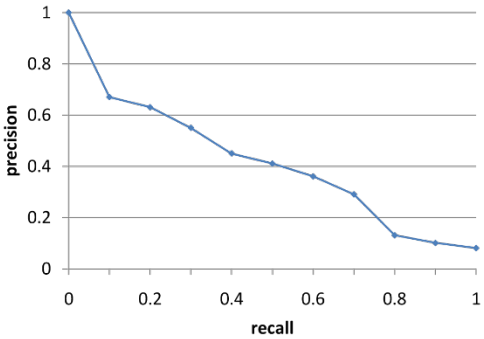

図 2・1　精度−再現率曲線の例

例えば，再現率レベルが（0.0, 0.1, 0.2, …, 1.0）であるときの補間精度を求めれば，**図 2・1** のような精度−再現率曲線を描くことができる．精度と再現率はトレードオフの関係にあるため，

精度-再現率曲線は一般には図 2･1 のような右下がりの形状となる．なお，上記の 11 点の再現率レベルに対する補間精度を平均したものを 11 点補間平均精度（eleven-point interporated average precision）という．

TREC においては，trec_eval というツール [*3] が用意されており，一定の形式でシステムの検索結果を与えれば，このツールにより精度-再現率曲線を描くためのデータが得られ，11 点補間平均精度（eleven-point interporated average precision）や，次節で説明する補間なし平均精度などの数値も計算される．

1-2-4 平均精度

テストコレクションを用いた評価によって検索システムの優劣を比較する場合，精度-再現率曲線よりも単一の数値によって比較が行えると都合がよい．先に示した 11 点補間平均精度もそのような指標であるが，実際には，補間なし平均精度（non-interpolated average precision）とも呼ばれる平均精度（average precision）が用いられることが多い．ある検索質問 q に対する平均精度$AP(Q)$ は，以下の式により算出される．

$$AP(Q) = \frac{\sum_{r=1}^{L} I(r)P(r)}{L} \qquad (2・2)$$

ここで，L は検索質問 Q に対して検索された文書の件数，$I(r)$ はランク r の文書が適合であるときに 1，不適合であるときに 0 となる関数，$P(r)$ はランク r までみたときの精度である．すなわち$AP(Q)$ は，ランクを 1 位から L 位へ向けて見ていったときに，適合である文書が得られるランクにおける精度をすべての適合文書にわたって平均したものである．また，$AP(Q)$ をテストコレクション中のすべての検索質問にわたって平均したものを全体平均精度（MAP : Mean Average Prevision）[*4] という．

1-2-5 その他の評価指標

ここまでに述べた評価指標は，いわば検索システムの優劣を比較するための評価指標であり，実際に検索システムを利用するユーザの視点とは必ずしも一致しないことがある．例えば Web 検索においては，そもそも適合文書の全体集合は明らかでなく，したがって再現率を議論することは無意味である．また，実際の Web 検索では，1 件または少数の適合するページが見つかればよいことが多い．このような場合の評価指標として，検索結果のランクの N 位までに含まれる適合文書の割合を表す P@N と呼ばれる指標がある．ここで N としては，Web サーチエンジンが最初のページに表示する件数（例えば 10 件）がとられることが多い．P@N は，Web 検索の利用実態を反映した指標であるといえるが，検索質問ごとに適合文書の件数が異なる場合には安定した評価指標とはいえないという問題がある．一方，とにかく 1 件の適合文書が見つかればよいという利用上の観点からすれば，最初の適合文書がランクの何番目に出現するかが評価指標になりうる．逆順位（reciprocal rank）は，このような評価指標であり，テストコレクションの全検索質問に渡す逆順位を平均したものを平均逆順位（MRR : Mean Reciprocal Rank）

[*3] http://trec.nist.gov/trec_eval/ から入手可能.

[*4] 平均精度，あるいは，MAP というとき，ある検索質問に対する $AP(Q)$ か，全検索質問にわたる平均なのかが混乱しないように注意する必要がある．

という．MRR は，3 章で述べられる質問応答タスクにおける評価指標として用いられている．

以上の評価指標はすべて，検索結果の文書が適合/不適合の二値に分類されるという前提によっているが，実際には検索された文書の有効性は連続的な性質を持つとも考えられる．このような性質に注目した評価指標として，例えば正規化減価累積利得（NDCG：Normalized Discounted Cumulative Gain）がある．NDCG は，文書の有効性を多値・連続的な利得ととらえ，各検索質問に対する検索結果のランクを第 1 位から L 位までみていったときの累積利得を全検索質問にわたって平均したものであるが，より詳しくは，参考文献 1）などの教科書を参照されたい．なお，質問応答や XML 検索などの新しい検索のタイプに適した評価指標については，解説記事（参考文献 2)）が参考となる．

■参考文献

1) C.D. Manning, P. Raghavan, and H.Shutze：Introduction to Information Retrieval, Cambridge University Press, 2008.
2) 酒井哲也：“よりよい検索システム実現のために．正解の良し悪しを考慮した情報検索評価の動向,”情報処理, vol.47, no.2, pp.147-158, 2006.

■7 群－6 編－1 章

1-3 テキスト情報検索の基本手法

（執筆者：林　良彦）［2015 年 5 月受領］

本節では，テキスト情報検索の基本手法について，特にユーザの検索質問（query）に対する適合度（relevancy）に応じて文書（document）をランキングする際に必要となる基本的な事項について説明する．

1-3-1 ベクトル空間モデル

(1)　ベクトルによる文書，検索質問の表現

将来的に自然言語理解技術が進展すれば，文書及び検索質問をその意味内容に基づいて表現し，両者の比較を行うことによる意味的な情報検索も可能となると考えられるが，現在までに用いられている代表的な手法は，ベクトル空間モデル（VSM：Vector Space Model）と呼ばれるもので，文書及び検索質問をこれらに含まれるターム（term）に基づくベクトルとして表現する．これらのベクトル表現を文書ベクトル，検索質問ベクトルと呼び，それぞれ，$\vec{D} = (w_{D_1}, \cdots, w_{D_T})$，$\vec{Q} = (w_{Q_1}, \cdots, w_{Q_T})$ のように書く．ここで，T はシステムのターム集合（vocabulary）の大きさであり，基本的なベクトル空間モデルにおけるベクトルの各要素は，システムのターム集合における一つのタームと対応する．また，w_{D_i} などで表されるベクトルの各要素をターム重み（term weight）という．当該のタームが文書/検索質問に出現しない場合は 0 を与え，出現する場合は後述するいずれかの手法により与える．VSM においては，文書や検索質問はシステムのタームが作る多次元空間における 1 点として表現され，構文的な特徴や意味内容はすべて捨象される．このような文書の表現を bag of words と呼ぶことがある．

(2)　タームの抽出

既に述べたように，タームとは単語，または，単語と同様に扱われる複合語や句を指す．検索対象の文書，また，検索質問から適切にタームを抽出する必要がある．タームの抽出処理としてどの程度の処理が必要かは，対象とする言語の特性や，所望のシステムの性質によって定まる．例えば，日本語のように単語の切れ目が自明でない言語の場合，単語を切り出す分かち書きの処理が必要となる．また，情報検索においては単語の正規化を行うことが多い．例えば，大文字・小文字を統一するといった簡単な処理もその一例であるが，これにより一般に再現率が向上する．更に，語形変化が生じている単語を基本形（lemma）に統一する（lemmatization），単語のつづりの不変化部分（stem）のみを残す（stemming）などの言語的な処理が行われる．単語より長い単位の複合語や句をタームとして認識する際は，専用の辞書を用いたり，構成要素の単語間の結びつきの強さを統計的に検定するといった処理が必要となる．一方で，英語における冠詞や前置詞，日本語における助詞などの機能語，対象文書において頻出する特殊な名詞などは，文書を特徴付けるものではないので，タームとして用いるのは適切ではない．このような語をストップワード（stop word）と呼ぶ．ストップワードは，品詞条件やあらかじめ用意したリストを用いて除去しておく場合がある．

(3)　ベクトルの類似度に基づく適合度の計算

VSM においては，文書ベクトルと検索質問ベクトルの類似度が検索質問に対する文書の適合度と対応すると考えるので，ある検索質問に対してある文書の適合度を計算することは，両

者のベクトルの類似度（similarity）を計算することに帰着される．したがって，情報検索システムがなすべきことは，与えられたユーザの検索質問に適合する文書を検索対象の文書集合から求め，それぞれのベクトルの類似度に基づいて適合文書を効率良くランキングすることとなる．

ベクトルの類似度の計算においては，両者におけるタームの分布が完全に一致している場合に最大値をとり，両者においてオーバーラップするタームが全く存在しない場合に最小値をとるように計算方法を設定する．このような要請を満たす計算方法の一つとして，ベクトルの内積（inner product）がある．すなわち，文書 D に対する検索質問 Q の適合度 $score(D,Q)$ を $score(D,Q)=\vec{D}\cdot\vec{Q}$ により求める．

ベクトルの類似度として内積を用いる場合は，多くのタームを含む長い文書の類似度が高くなる可能性がある．これを避けるためには，文書長に関する正規化を行うことが必要となる．このような方法の一つとして，両者のベクトルが成す角θ_{DQ} のコサイン値に基づく方法（cosine measure）がよく用いられる．すなわち，文書 D に対する検索質問 Q の適合度$score(D,Q)$ を次式により求める．

$$score(D,Q)=\cos\theta_{DQ}=\frac{\vec{D}\cdot\vec{Q}}{[\vec{D}][\vec{Q}]} \qquad (3\cdot 1)$$

上記の計算式において，両者のベクトルが直交している場合は分子の値（2 つのベクトルの内積）は 0 となり，適合度は 0 となる．これは，検索質問と文書の間にオーバーラップするタームが存在しない場合である．一方，両者が完全に一致している場合の適合度は 1 となる．すなわち，コサイン値に基づく適合度は 0 から 1 の範囲に正規化されている．

後でも述べるように，文書長をどのように扱うかは適合度計算における一つのポイントである．上記のコサイン値を用いる場合，内積を用いる場合とは逆に，直観に比べて短い文書の適合度が高くなりやすい．ピボット文書長正規化（pivoted document length normalization）は，関連度判断が付与されたテストコレクションを用いて，文書長の影響を実験的に調整する方法である．

(4) tf-idf 法によるターム重み付け

上記では文書ベクトルの各要素のターム重みを w_{D_i} のように書いた．w_{D_i} の与え方（term weighting）として最も単純な方法は，当該のタームが文書 D に出現すれば 1，出現しなければ 0 とする方法であるが，そのタームが文書 D を特徴付ける度合に応じた重みを与えることが望ましい．

以下では，tf-idf 法と呼ばれる手法を説明する．この手法は，(1) あるタームがある文書中で多く用いられていればその文書を特徴付けるタームである可能性が高いが，(2) ただし，多くの文書に現れるようなタームであれば，その文書を特徴付けるものにはならない，というヒューリスティックに基づく．

次式は，最も基本的な tf-idf 法[*5] によるターム重みの計算式である．

$$w_{D_i}=tf_{D_i}\times\log\frac{N}{df_i} \qquad (3\cdot 2)$$

[*5] tf-idf 法の計算式には様々なバリエーションが存在する．それらについては，文献 1) を参照されたい．

ここで，第 1 項をターム頻度（term frequency）項と呼ぶ．tf_{D_i} は，ターム i の文書 D における頻度を表す．また，第 2 項を逆文書頻度（inverse document frequency）項と呼ぶ．df_i は，検索対象の文書集合においてターム i を含む文書の数（document frequency）であり，N は検索対象の文書集合の大きさ，すなわち，全文書数である．第 1 項が上記の仮定(1)，第 2 項が上記の仮定(2)にそれぞれ対応することは明らかであろう．実際，あるタームが文書集合中のすべての文書に現れるものであれば，N/df_i が 1 となることから第 2 項の値は 0 となりターム重みも 0 となる．一方，特定の文書にしか現れないようなタームについては N/df_i の値が大きくなるが，その影響を適当に減衰させるため log をとる．

検索質問に対する適合度の計算を行うためには，検索質問中のターム重み w_{Q_i} も適切に与える必要がある．これには，単純に検索質問中の各タームの出現を 1/0 で表す方法，検索質問中の各タームの出現頻度を用いる方法のほか，文書の場合と同様に tf-idf 法に基づく方法などのバリエーションがある．

1-3-2 再考：適合度の計算

情報検索システムのなすべき仕事は，与えられたユーザからの検索質問に適合する文書を効率良く適合度順にソートして提示することである．特に Web 検索のように文書集合が膨大な場合は計算の効率が大きな問題となる．この観点から，コサイン値に基づく適合度の式を見直してみると，分母における検索質問ベクトルの大きさの項は，適合度の大小に影響しないので，検索結果文書をランキングするという目的においては計算する必要はない．また，適合度の計算においては，検索質問中に含まれないタームは文書中に出現していたとしても適合度の計算には関与しない．以上から，適合度の計算式は次式のように書くことができる．ここで，検索質問を $Q' = (q_1, \cdots, q_j, \cdots, q_n)$ と表し，q_j を n 個のタームからなる検索質問中の j 番目のタームとしている．

$$score(D, Q') = \sum_{j=1}^{n} W_Q(q_j) \times W_D(q_j, D) \qquad (3・3)$$

上式において，$W_Q(q_j)$ は検索質問中のターム q_j の重みを表し，先に示したいずれかの方法による重みを与える．また，$W_D(q_j, D)$ は文書 D におけるターム q_j の重みを表し，$TF(q_j, D)$ を文書 D におけるターム q_j の頻度，$L(D)$ を文書 D の文書長，ターム q_j の文書頻度を $DF(q_j)$ とするとき，次式で表される．もちろん，右辺の第 1 項はターム頻度項，第 2 項は逆文書頻度項に対応する．

$$W_D(q_j, D) = \frac{TF(q_j, D)}{L(D)} \times \log \frac{N}{DF(q_j)} \qquad (3・4)$$

検索質問中のターム重みを一様に 1 とし，逆頻度項を $IDF(q_j)$ と書くと，適合度 $score(D, Q')$ の計算式は以下のように書ける．

$$score(D, Q') = \sum_{j=1}^{n} IDF(q_j) \times \frac{TF(q_j, D)}{L(D)} \qquad (3・5)$$

1-3-3 OKAPI BM25

OKAPI BM25（または，単に BM25）は，TREC においてその有効性が確認されている手法であり，現在の Web サーチエンジンにおいても利用されている手法である．ここでは詳しく述べないが，BM25 は確率的情報検索（probabilistic information retrieval）と呼ばれる情報検索のモデルに基づいて理論的に導出される．BM25 は BM25 ターム重み（term weighting）と呼ばれることもあるが，実際には次式のような適合度 *6 の計算式として示される．ここで，L_{ave} は文書集合における平均文書長である．

$$score(D,Q') = \sum_{j=1}^{n} IDF'(q_j) \times \frac{(k_1+1) \times TF(q_j,D)}{k_1((1-b)+b \times (L(D)/L_{ave})) + TF(q_j,D)} \quad (3・6)$$

上記で，k_1 はターム頻度の影響を調整する非負のパラメータであり，$k_1 = 0$ とすると適合度は逆文書頻度項のみにより決まることになる．一方，b は文書長の影響を調整する 0 以上 1 以下の値をとるパラメータであり，$b = 0$ とすると文書長の正規化を行わないことになる．実際には，$k_1 = 2$，$b = 0.75$ という値がよく用いられる．

$IDF'(q_j)$ としては，通常の逆文書頻度 $IDF(q_j)$ を用いてもかまわないが，BM25 においては確率論的な裏付けから，ターム q_j を含む文書数と含まない文書数の比として特に次式のように定義される．ここで，分母・分子に加算されている 0.5 は，スムージングのための定数である．また，N が $DF(q_j)$ に比べて十分大きければ，$IDF'(q_j) \approx IDF(q_j)$ となることを注意しておく．

$$IDF'(q_j) = \log \frac{N - DF'(q_j) + 0.5}{DF'(q_j) + 0.5} \quad (3・7)$$

BM25 では，ターム q_j が文書集合中の半数以上の文書に出現するようなポピュラーなタームである場合，上記の $IDF'(q_j)$ は負の値をとることに注意しておく必要がある．これを避けるためには，あらかじめ文書頻度の大きい，すなわち，ストップワード的な性格の強いタームをシステムのターム集合から除いておくという処理が必要である．最後に，VSM におけるコサイン値に基づく適合度と BM25 による適合度は理論的な背景は異なるが，式(3・4)と式(3・5)は類似した形となっていることを注意しておく．

1-3-4 適合性フィードバック

実際の情報検索においては，一度の検索質問で所望の文書が得られるとは限らない．この場合，検索質問を修正して再度検索を行うことになるが，一度目の検索結果から得られる情報をうまく利用して適切に検索質問を修正できればよい．適合性フィードバック（relevance feedback）とは，初期の検索結果に対して与えられた適合度判定を用いて検索質問ベクトルを更新する方法である．次式は，Rocchio の式と呼ばれる検索質問ベクトルの更新式である．ここで，D_R は適合文書の集合，D_N は不適合文書の集合である．要は，適当な重み α, β, γ のもとで，前回の検索質問ベクトル $\overrightarrow{Q_0}$ に適合文書におけるターム重みを加え，かつ，不適合文書からのターム重みを減算することにより修正された検索質問ベクトル $\overrightarrow{Q_m}$ というもので，直感的にわかりや

*6 確率的情報検索における適合度は検索状態値（RSV：Retrieval Status Value）と呼ばれることがある．

すい式となっている．適合性フィードバックは，原理的には精度も再現率も向上させるが，実際には再現率を改善することに効果があると言われている．

$$\overrightarrow{Q_m} = \alpha\overrightarrow{Q_0} + \beta\frac{\sum_{D_i\in D_R}\overrightarrow{D_i}}{|D_R|} - \gamma\frac{\sum_{D_j\in D_N}\overrightarrow{D_j}}{|D_N|} \qquad (3・8)$$

実際の運用においては，検索結果中のすべての文書に適合度判定を付与するのは困難なので，適当な数の目についた適合文書のみをユーザに選ばせ，これを用いることも多い．この場合 $\gamma = 0$ とするポジティブフィードバックを行うことになる．適合性フィードバックの機能は，実際の Web サーチエンジンでも“More like this”といった名称で実装されたが，あまり用いられることはなかった．一方で，初期の検索結果における一定数の上位の結果をユーザの判断を仰ぐことなく適合文書とみなしフィードバック検索を行う疑似適合性フィードバック（pseudo relevance feedback）が通常のアドホック検索や言語横断検索の性能を改善するのに有効であることが TREC において明らかとなっている．

1-3-5 その他の情報検索モデル

既に述べたように BM25 は，確率的情報検索モデルに基づき導出される．これまでには，ベクトル空間モデルや確率的モデルのほかにもいくつかの情報検索のモデルが提案されており，最近では特に言語モデル（language model）に基づく情報検索が注目を集めている．言語モデルに基づく情報検索の研究により，これまでヒューリスティックと考えられてきた tf-idf 法に理論的な根拠があることが明らかとなった．

また，基本的なベクトル空間モデルの問題点を解決する方法として，潜在的意味インデクシング（LSI：Latent Semantic Indexing）と呼ばれる方法が提案されている．すなわち，ベクトル空間モデルにおいては，システムの語彙中の各タームを独立した次元と考えていたため，タームの形としては同じだが意味が異なる用法が同一次元となってしまう，逆に，異なるタームが同じ意味を表していたとしても異なる次元として扱われるという問題があった．LSI では，各タームの各文書における出現を表すターム・文書行列（term-document matrix）を特異値分解（SVD：Singular Value Decomposition）という数学的手法により次元圧縮することにより，タームの重み付き線形和により表現される新たな次元が導かれる．これらの各次元のそれぞれは潜在的な（latent）意味を表すものと考えられ，このような次元を用いることにより，上記の 2 つの問題の解決が図られるとされている．LSI では文書も検索質問も次元圧縮された新たな空間におけるベクトルとして表現されるため，基本的なベクトル空間法の拡張であると考えることができる．

■参考文献

1) C.D. Manning, P. Raghavan, and H.Shütze：Introduction to Information Retrieval, Cambridge University Press, 2008.