2N5-OS-21b-1

sui-sei：消費者インサイトをリアルタイムに捉えるための EC サイト向け高速検索エンジンサービス

sui-sei: Site Search Engine for E-Commerce

山本 浩生[*1]
Kose Yamamoto

岩瀬 高博[*1]
Takahiro Iwase

[*1] 株式会社 神戸デジタル・ラボ
Kobe Digital Labo, Inc.

We developed a site search engine for E-Commerce, named “sui-sei”. To fast-search for commodities in an E-Commerce site, sui-sei uses NOSQL database “okuyama”. For a site search, not only a speed of searching, but also gaining consumer insight is required. In this paper, we explain the architecture of sui-sei, and show a result of verifications of search speed. And we consider how to gain consumer insight.

1. はじめに

近年、商品販売を目的とした E-Commerce サイト(以下、「EC サイト」とする)市場規模が拡大するにつれ、サイト内検索エンジンサービス市場も拡大を続けている。[矢野研究所 2011]また、サイト内検索エンジンは単に与えられた検索語に適合する結果を返却するだけではなく、可能な限り多くの商品を販売したいというビジネスの要求を実現するため EC サイトに特化した検索機能を備えるようになった。大量の商品の中から高速に検索を行う基本的な機能に加えて、検索対象から予測される関連商品情報を検索結果に含めて提示するレコメンドと呼ばれる機能などがこれにあたる。

以上の状況を踏まえて、株式会社 神戸デジタル・ラボ(以下、「KDL」とする)では、EC サイト内検索エンジンサービス sui-sei [sui-sei] を開発、運営している。sui-sei の特徴は、検索を行うためにリレーショナルデータベースを用いず、NOSQL データベースである okuyama [okuyama] を利用することで検索速度性能と安定性の向上を図っている点である。本稿では、サイト内検索エンジンである sui-sei を概説し、クラウド環境で NOSQL データベースを採用したサイト内検索エンジンが実用に足る性能を有することを説明するとともに、サイト内検索エンジンサービスを提供することで得られた知見を基にビジネスから見た検索エンジンの現在の課題と今後のサイト内検索エンジンに求められる要素を考察する。

2. sui-sei

2.1 sui-sei の構成

sui-sei は検索インデックス及び検索データを格納する NOSQL データベースである okuyama クラスタと、検索 web インターフェースアプリケーションから構成される。NOSQL データベースである okuyama の持つデータの大量保存、高速なデータ処理といった特徴[岩瀬 2012]を検索エンジンとして活用する。運用中の sui-sei はサーバー3 台で構成されている。全 3 台で 1 つの okuyama クラスタを構築し、それぞれのサーバーに検索 web インターフェースアプリケーションが稼動しており、ロードバランサーで web インターフェースアプリケーションへのアクセスを振り分けている。また、特定の 1 台にのみ検索データのインポートプログラムがインストールされている。(図 1)

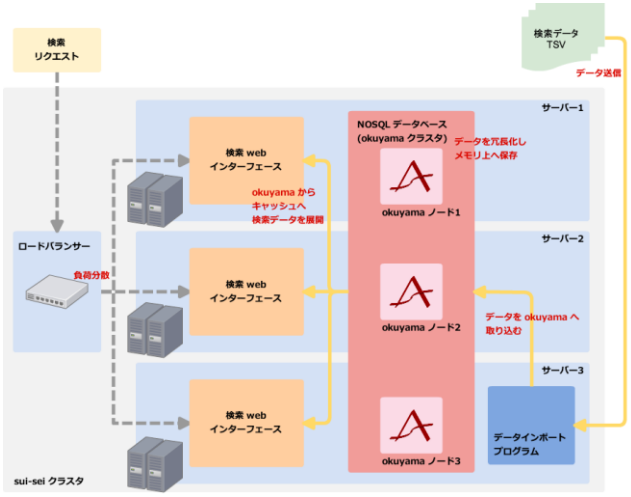


図 1：sui-sei の構成

sui-sei の処理は下記の順で行われる。

(1) 検索対象データの取り込み

検索用データと検索インデックスを作成するためのインポートプログラムが TSV 形式の検索データファイルを okuyama クラスタに検索用データと検索インデックスとして登録する。

(2) 検索条件の取得

検索 web インターフェースが検索リクエストの URL クエリパラメータから検索条件を取得する。

(3) 検索キャッシュのデータ探索

検索条件に応じた検索結果データをメモリキャッシュから探索してデータが存在すれば利用する。

(4) NOSQL データベースからのデータ探索

メモリキャッシュにデータが存在しない場合は okuyama クラスタからデータを取捨選択し検索結果を作成する。

(5) 検索結果の返却

リクエスト元に検索結果データを返却する。

連絡先：山本 浩生，株式会社 神戸デジタル・ラボ，ko-yamamoto@kdl.co.jp

なお、メモリキャッシュへの検索結果登録は検索完了後に加えて、定期的に実行されるメモリキャッシュ登録用のワーカーによって行われる。

2.2 性能検証

(1) 要求性能と検証条件

sui-sei に要求する検索性能と検証の条件を次の通り設定した。

- それぞれ約 40 字から約 100 文字程度の商品説明項目を持つ、商品 5 万件のデータを NOSQL データベースに登録しておく
- 1 秒間に 70 回のリクエストを実行
- 各リクエストに対するレスポンスが 1 秒以下

これは大規模 EC サイトで負荷が著しく増加する EC サイト繁忙期の検索実績から算出した数値である。

検証環境としてはクラウドコンピューティングサービスを利用することとし 3 社を選定した。検証に利用したクラウドコンピューティングサービスのインスタンスは 3 社ともにほぼ同程度の性能のプランを選択し、それぞれが 4 コア CPU 、メモリ 16GB を搭載している。

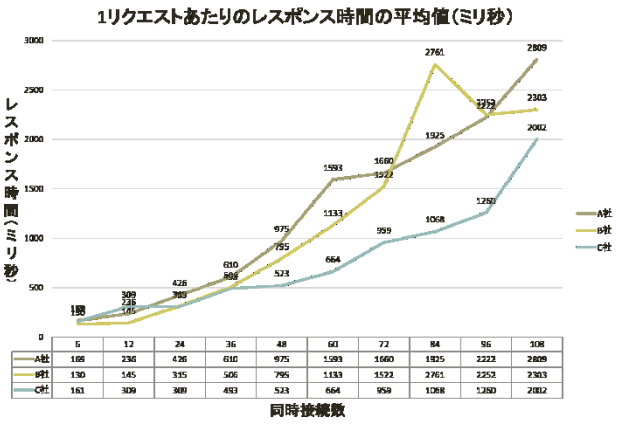
(2) 検索方法

サイト内検索には商品名検索やカテゴリ検索などがあるが、検証では商品説明に対する全文検索を行うこととし、検索リクエストは sui-sei を動作させるネットワークの外部に置かれたサーバーから Apache JMeter [JMeter] を利用して生成する。テストシナリオとして、1 秒間の検索回数が 6～108 回となるように 2 分間隔で秒間検索回数を 12 回ずつ増加させ、それぞれ 1 回毎の検索結果商品件数が 100 件から 4992 件までの 12 段階の中から無作為に選ばれるようにして計測した。

また長時間負荷を与え続ける検証として、秒間リクエスト数を 48 回に固定してそれ以外の条件は変更せずに 6 時間リクエストを続け、検索性能に劣化が生じないか検証した。これは繁忙期において特に連続して検索が高負荷となる時間が 3 時間であること、その後の 3 時間も通常時期に比べて高負荷となることから合わせて 6 時間と設定している。

2.3 結果・考察

検証結果を図 2 に示す。



	5	12	24	36	48	60	72	84	96	108
A社	168	236	426	610	975	1593	1660	1925	2223	2809
B社	130	145	315	506	795	1133	1522	2761	2252	2303
C社	161	309	369	493	523	664	959	1068	1260	2002

図 2：1 リクエストあたりのレスポンス時間の平均値(ミリ秒)

利用したクラウドコンピューティングサービスにより検索性能の差が発生しており、要求性能として設定した値である 70 並列リクエストに対して 1 秒以下のレスポンスを達成したのは図の C 社のみであった。これはクラウドコンピューティングサービスが公表している性能が同程度であっても A 社 B 社環境のネットワーク性能と CPU 性能が C 社のものと比較して低いことによる影響と考えられる。しかしながら、A 社 B 社の場合であっても今回検証に利用したものよりも上位性能のインスタンスを用いることで要求性能に達すると考えられる。実際にサービス化する場合には C 社環境を選択することから、要求性能を満たしているとみなせる。

また、長時間負荷を与え続ける検証の結果は図 3 のようになった。レスポンス時間の変化は見られないことから安定した性能を有することがわかった。これは NOSQL データベースを利用したことで、データの取得速度が一定となった結果によると推測される。

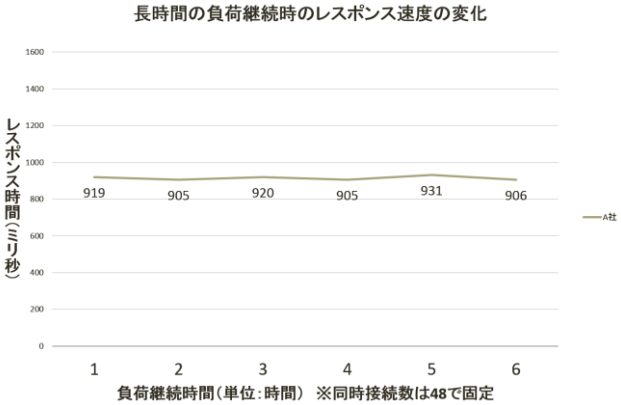


図 3：長時間の負荷継続時のレスポンス速度の変化

以上の結果から、クラウド環境で NOSQL を利用したサイト内検索エンジンは速度面で実用に足るとわかった。

2.4 課題

検証では検索速度の面からサイト内検索サービスが実用可能であるかを判断したが、実際には検索速度だけでサイト内検索エンジンの性能を測ることはできない。これはサイト内検索が EC サイト内で商品の検索に利用されるため、検索速度に加えて消費者インサイトを捉える必要があることによる。次章では消費者インサイトを仮定した上で、消費者インサイトを捉えるサイト内検索エンジンについて検討する。

3. サイト内検索エンジンと消費者インサイト

3.1 サイト内検索エンジンの特性

サイト内検索エンジンと Web ページ検索に一般的に利用される検索エンジンには差異が存在する。Web ページ検索では不特定多数の中から検索を行うため検索結果も多くなるとともに、検索結果は検索条件である単語や文字に対して正確なものとなりやすい。一方でサイト内検索エンジンは特定の商品の中から最も望ましい商品を検索結果として提示するため検索結果数がより絞り込まれたものとなり、検索結果は検索条件そのものではなく条件から判断される商品や消費者のニーズに対して正確なものとなるべきである。

具体例として「バック」という検索条件語句を考えた場合、Web ページ検索エンジンではまず優先的に「バック(back)」の語それ自体の検索結果を返却し、次に検索条件のゆらぎを考慮して「バッグ(bag)」や「パック(pack)」の検索結果を返却する。しかしサイト内検索エンジンでは取り扱っている商品に近似した結果が正確であるとするため、「バック」ではなく「バッグ」の検索結

果が消費者の必要としている情報であると判断してバッグの検索結果を提示することが望まれる。この場合商品として「バック」が存在しない場合は「バック」で検索が行われ検索結果が 0 件となり、表示できる商品が存在せず販売機会の損失が発生している。

よって、サイト内検索エンジンには消費者の潜在的ニーズ、つまり消費者インサイトを捉える検索結果を提示する必要性が存在すると言える。

3.2 消費者インサイト

検索を行った意図通りに検索結果が返却されることは検索エンジンとして必要な機能であるが、サイト内検索エンジンには検索を行った消費者自身にすら無自覚なニーズを喚起して商品の購入を促すことも必要とされる。この無自覚な消費者ニーズである消費者インサイトを捉えるために、検索結果を要求された条件に完全一致する商品だけではなく、条件に一致する商品に類似している商品や趣向の似ている消費者が求めた商品も含めて提示する。

3.3 消費者インサイトを捉えるサイト内検索手法

消費者インサイトを捉えたサイト内検索エンジンを実現するための方法例を挙げる。

(1) ユーザーの検索履歴に応じて最適化された検索

ユーザー自身の検索履歴を蓄積したデータを基にして検索結果にユーザーの趣味趣向を反映させることで、検索条件に対応する商品を最もユーザーの必要としている商品に近づける方法である。この方法は商品の選別と表示順序で利用可能である。

(2) 人気商品を反映した検索

ユーザー個人の趣味趣向とは無関係に全ユーザーの利用履歴から判断した人気商品を検索結果に加えて提示する方法である。

(3) 0 件ヒット問題に対応した検索

利用者の入力した検索条件に適合する検索結果が存在しない場合に、利用者が必要としていると想定される検索結果を作成して提示する方法である。

以上の例(1)は消費者が行う検索に対して検索条件に最適な検索結果を提示するという検索の最適化の性質が強く、消費者が検索前に購入を検討していた商品の購入をスムーズにする効果が予想される。一方で、例(2) (3)はいずれも検索結果に消費者が購入を検討しやすい商品の選択肢を追加して提示している。例(2)(3)は検索時に具体的に購入を検討していなかった商品の購入に繋がる可能性があるため、消費者の潜在ニーズを喚起する、つまり消費者インサイトをより一層捉える性質があると考えられる。

また、課題として次の内容が考えられる。例(1)は個人情報の問題を含んでいる。具体的には、ユーザーの検索履歴やサイトの利用履歴を収集して利用することから、事前にユーザーへ行動履歴を収集していることを明記してオプトイン型とすることが望ましい。例(2)では、全体で購入件数の多い商品が必ずしもそのユーザーにとって望ましい商品であるとは限らない点に課題が残る。この課題を解決するため、例(2)は例(1)の技術と併用することで検索精度が向上すると思われる。

以上の検索方法の中から、sui-sei では消費者インサイトを捉える検索として 0 件ヒット問題を取り上げて対策を行った。

4. 0 件ヒット問題への対応

ユーザーの入力した検索条件に対して有効な結果を持たず検索結果が 0 件となる場合、EC サイト事業者としては販売機会を損失している。このような販売機会の損失を防ぐため、検索結果が 0 件であった際に類似の検索条件を作成し、何らかの検索結果を提示することが 0 件ヒット問題への対応策となる。

4.1 0 件ヒット問題の発生割合

0 件ヒット問題への対応を行う前に、0 件ヒット問題の発生割合について調査を行った。sui-sei を実際のサービスに導入することで得られた 150 万回の検索履歴の中からキーワード検索の行われた件数とその検索結果が 0 件であった件数を計算すると、約 9.1%のキーワード検索で検索結果が存在しない 0 件ヒット問題が発生していた。

4.2 0 件ヒット問題への対応方法

sui-sei では 0 件ヒット問題に対応する方法として、株式会社きざしカンパニーの提供する Kizasi Interface for Syndication Services (KISS) [kizasi]を利用した。きざしカンパニー社はブログの投稿内容データを解析して特定の語句と語句との関連性を解析した結果をデータベース化しており、KISS は語句を与えることでそのデータベースの中から関連性の高い語句(共起語)を API 形式で取得するものである。

まず sui-sei は検索処理を開始する。検索条件に一致する商品が存在した場合は通常通り検索結果とする。検索条件に一致する商品が存在しない場合には、検索条件の中から検索語句を取り出し KISS に与えて共起語を取得する。KISS から取得した共起語を検索語句に設定し、当該箇所以外は元々の検索条件を利用して再度 sui-sei に対して検索リクエストを行う。2 度目の検索で結果を得ることが可能であれば、0 件ヒット問題への対応が成功している。なお、2 度目の sui-sei への検索リクエストで検索条件に一致する商品が存在しない場合は検索結果が 0 件となる。

4.3 対応効果検証

0 件ヒット問題対策の効果検証として、実際に 0 件ヒット問題の発生している検索語句を取得し、検索語句について KISS を使用した結果得られる共起語で再検索を行った場合に、検索結果が 1 件でも存在しているか調査を行った。

(1) 検証用語句抽出と検証方法

キーワード検索を行った検索結果件数が 0 件となった検索キーワードを、検索に使用された回数で順位付けし、上位 15 語を検証に用いた。1 つの元検索語句から得られる複数の共起語の中から、共起性の高い順に名詞を 3 語選んで新検索語とし、sui-sei で再検索し結果を確認した。

(2) 検証結果

15 の検索語句の内、9 語で検索結果が存在した。この内 2 件は再検索で得られた検索結果が元検索語と意味的に強く結びついた商品を表示することができた。一方で、検索結果は得られたものの、表示された商品と元検索語句に全く関連の見られないものが 4 語存在した。また 6 語では共起語を取得できなかった。よって 15 の検索語句中、5 語で 0 件ヒット問題に対応できていたが、残りの 10 語で対応することができていないことがわかった。(図 4)

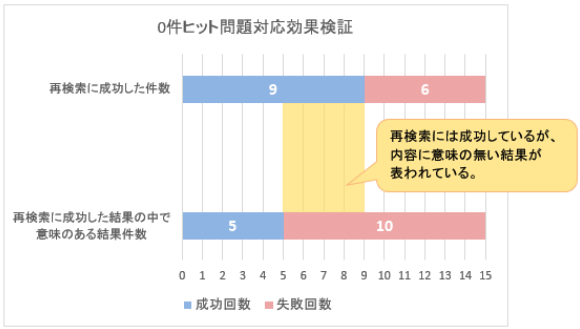


図 4：0 件ヒット問題対応結果

(3) 課題

共起語の取得が行えなかった検索語句が存在することから、今後は共起語取得に用いる辞書の拡充が課題である。また、共起語を利用した再検索の結果が元の検索語句を考慮すると意味を成していない結果が 4 件存在したため、検索の文脈や EC サイトの必要とする語句のカテゴリや分類に応じて共起語として取得する語句を限定し、再検索の結果が元の検索語句に意味的に近いものとなるように処理することも課題として挙げられる。

このため今後は、検索対象の商品に応じた辞書を用意するなど共起語の取得に用いる辞書の拡充を目指すとともに、共起語として取得する語句を必要に応じて取捨選択する処理を追加することが必要である。

5. まとめ

本稿では我々の開発したサイト内検索エンジンである sui-sei を基に、クラウドコンピューティング環境で NOSQL データベースを採用した検索エンジンが実用に足る性能を発揮可能であることを示すとともに、サイト内検索エンジン特有の性質である消費者インサイトを捉える必要性とその具体的な方法について述べ、その内 0 件ヒット問題への対応を行って効果を検証した。今後は 0 件ヒット問題への対応策の改善と効果の再検証を行うとともに、例として挙げた消費者インサイトを捉えるサイト内検索手法について検証を進めていきたい。

参考文献

[矢野研究所 2011] 矢野研究所: エンタープライズサーチ・サイト内検索エンジン市場に関する調査結果 2011, 矢野研究所, 2011.

[岩瀬 2012] 岩瀬他: ビッグデータ・イン・メモリ, 2012 年度人工知能学会全国大会, 1A1-OS-17a-3, 2012.

[sui-sei] sui-sei 公式サイト http://sui-sei.jp/

[okuyama] okuyama 公式サイト http://okuyama-project.com/

[JMeter] Apache JMeter http://jmeter.apache.org/

[kizasi] 株式会社きざしカンパニー http://kizasi.jp