
Namazu を PHP と FLASH で調理しましょう

～グループ研究会の活動を通して～

FUJITSU ファミリ会関西支部 平成 16 年度

IT フォーラム「Linux 研究会」B グループ

■ 執筆者 Profile ■



中野 成年

2002 年 (株) ダイテック 入社
システム業務担当

2005 年 現在 営業グループ主任
システム・営業担当



外村 育博

1992 年 (株) たけでん入社
社内システムの開発・教育
を担当

2005 年 現在 情報システム部課長代理
主に社内ポータルサイトの
開発業務を担当



高土 俊輔

2003 年 日本ピラー工業 (株) 入社
システム業務担当

2005 年 現在 情報システム部所属
システム担当



橋本 英晃

1996年 福西電機 (株) 入社
エンジニアリング総合本部
システム課
2005年 現在 エンジニアリング総合本部
システム課所属 営業担当



廣瀬 学

2000年 中央コンピューター (株) 入社
システム開発業務担当
2005年 現在 大阪事業本部
ビジネスソリューション1部
2グループ所属
Web システム担当



高橋 弘和

2002年 上村工業 (株) 入社
情報本部情報管理部勤務
業務システムの開発,
ホスト系ハードウェアの管理担当
2005年 現在 同上

■ 論文要旨 ■

現在, IT ソリューションは, 一昔前とは異なり, Linux などオープンソースとの組み合わせによる様々な提案が行われている。

平成 16 年度 FUJITSU ファミリー会関西支部『IT 研究フォーラム』では Linux と Apache, Namazu (検索エンジン) を素材とし, Linux によるシステム構築をテーマに実施され, 私達Bグループがそのテーマに対して, どのように取り組み, どのような結果を得られたかを改めて振りかえることにした。

テーマのとおり, サーバ環境の OS には Linux, WEB 環境に Apache, 検索に Namazu, 検索キー入力画面に FLASH (フラッシュ), 結果表示に PHP (プログラム言語) を使いユーザーにいかに使ってもらえるシステムができるか, それを実現するためにどのような問題が潜み, 解決していかなければならないかを7ヶ月間の活動を参考に, 今後活かして頂ける様に, 論文として纏めた。

■ 論文目次 ■

| | |
|---------------------------------|-------|
| 1. はじめに | 《 4》 |
| 1. 1 開発に至った経緯 | |
| 2. 定着までの問題点 | 《 4》 |
| 2. 1 検索ソフトが利用者に定着しにくい | |
| 2. 2 情報を常に最新に保つのは運用者の負担 | |
| 3. 開発の目的(コンセプト) | 《 5》 |
| 3. 1 定着させる為に | |
| 3.1.1 誰にでも分かりやすいシステム | |
| 3.1.2 誰にでも管理できるシステム | |
| 4. システムについて | 《 9》 |
| 4. 1 システム構成 | |
| 4. 2 プログラム間連携フロー | |
| 4. 3 Wgetパラメータ設定システムによる情報収集イメージ | |
| 4. 4 INDEX 作成のイメージ | |
| 4. 5 利用者が検索したときのイメージ | |
| 5. 開発における問題点 | 《 12》 |
| 5. 1 Namazu検索結果の文字化け | |
| 5. 2 検索結果がすべて0件で返される | |
| 5. 3 見えないエラー | |
| 6. システムの評価 | 《 13》 |
| 7. 今後の課題 | 《 13》 |
| 8. おわりに | 《 13》 |

■ 図表一覧 ■

| | |
|---|-------|
| 図1 『森田検索』操作画面<トップ画面>イメージ図 | 《 5》 |
| 図2 『森田検索』操作画面<AND検索>イメージ図 | 《 6》 |
| 図3 『森田検索』操作画面<OR検索>イメージ図 | 《 6》 |
| 図4 『森田検索』操作画面<NOT検索>イメージ図 | 《 7》 |
| 図5 『wgetパラメータ管理システム』画面図 | 《 7》 |
| 図6 プログラム間連携フロー図 | 《 9》 |
| 図7 wgetパラメータ設定システムによる情報収集イメージ図 | 《 10》 |
| 図8 INDEX作成のイメージ図 | 《 11》 |
| 図9 利用者が検索したときのイメージ図 | 《 12》 |

1. はじめに

1. 1 開発に至った経緯

私達Bグループは、FUJITSU ファミリー会『IT 研究フォーラム』において、様々な問題を解決し成果を出すことに成功した。これらに至った経緯も含め以降にまとめている。

まず、『IT 研究フォーラム』での研究において何を成果とすべきか検討を行った。今回のテーマである、Linux, Apache, Namazu の中から、まず Namazu の性能について比較検討することとした。その他、商用検索ソフトの機能比較や他のオープンソースとして提供されている検索ソフトなどの評価も行う事を検討した。

検討の結果、各社にとっても有意義な成果を出さなければならないという課題に立ち返り、自社に戻った後、日常業務に役立てられる成果物を開発目標とした。

そこで、自社で検索ソフトをどのような事で活用すれば効果があるかを話し合った。

その中で業務に役立ち、定着させるには、使い方が分かりやすく、すぐに使える事がメンバーからの意見として挙がった。

確かに、検索ソフトは便利であるが、目的の情報が素早く的確に見付からなくては定着にはつながらない。

これを解決する事が第一の目標となった。次に管理者が管理しやすく、日々の業務の中で負担にならず、データを最新に保てる運用ができることが第二の目標として挙がった。

親しみのあるシステムという観点から、コンテンツで幅広く利用されている FLASH を思い切って採用し、視覚的に検索目的がはっきり分かる画面作りを目標とした。この開発で、FLASH と PHP, Namazu を連携させるパターンは意外な組み合わせであるが、私達のグループ内のスキル、開発効率や目標達成の両面から最適という判断となった。

システム名は「森田検索」とした。

2. 定着までの問題点

2. 1 検索ソフトが利用者に定着しにくい

検討をすすめる中で、システムを導入しても定着させなければ意味がなく、どうやって利用しやすいシステムを構築するかが焦点となった。

考えられる要因としては、キーワード入力機能が以外に活用されていない事である。

検索ソフトでは、AND, OR, NOT 等々を組み合わせで絞り込んでいかなければ、本来の効果を発揮しないのである。

また、定着しにくいという背景には、単一キーワードで検索し、沢山の検索結果から更に目で判断しなければならないという苦勞が考えられる。検索システムを使い慣れないユーザーには理解しにくい部分である。使い慣れたユーザーで有っても更に高度な機能を使いこなすにも高度な組み合わせで検索するのは手間なものである。

これらの問題を解決するために利用者が検索ソフトの機能を理解する必要が無く、AND, OR, NOT, グループ化, 部分一致, 正規表現検索, フレーズ検索, フィールド指定, NEAR 検索, 重み指定, 類似検索などの組み合わせで絞り込み検索が容易にできることが重要であると判断した。

2. 2 情報を常に最新に保つのは運用者の負担

この課題は、検索システムの利用者が定着しても検索結果が陳腐化、新たなジャンルの情報が追加されないなど、情報が最新で無くなると利用者が離れてしまう可能性がある。それを防ぐには、インデックスデータを最新に保つ必要がある。そのためには、Namazu が持つデータを収集しインデックス化する機能をコマンドにより実行しなくてはならない。結果として、Namazu に関する基本的な知識を有することも必要となり、運用管理者の負担になると考えられる。

3. 開発の目的(コンセプト)

3. 1 定着させる為に

先に述べた問題点を克服するために、私達は以降に挙げる解決策をシステムに盛り込むことにした。

3. 1. 1 誰にでも分かりやすいシステム

検索キーワード入力画面をビジュアル化（視覚化）し、入力した検索キーワードがどのような範囲で検索されるかを表現できる画面機能を付加することにした。このことにより利用者は、AND, OR, NOT, グループ化, 部分一致, 正規表現検索, フレーズ検索, フィールド指定, NEAR 検索, 重み指定, 類似検索などの機能を理解することなく検索範囲を把握できる。今回の開発では、すべての検索に対応することは時間的に難しい為、AND, OR, NOT に絞り最も基本的な部分で基盤となるシステムを目指した。

以下に画面イメージを紹介し、どのような解決策を見出したか紹介する。

3. 1. 1. 1 「森田検索」のトップイメージ

まず FLASH で作成したトップ画面が表示される。マスコットを Namazu を利用した検索システムであるイメージが伝わるよう工夫し、より親しみあるものとした。

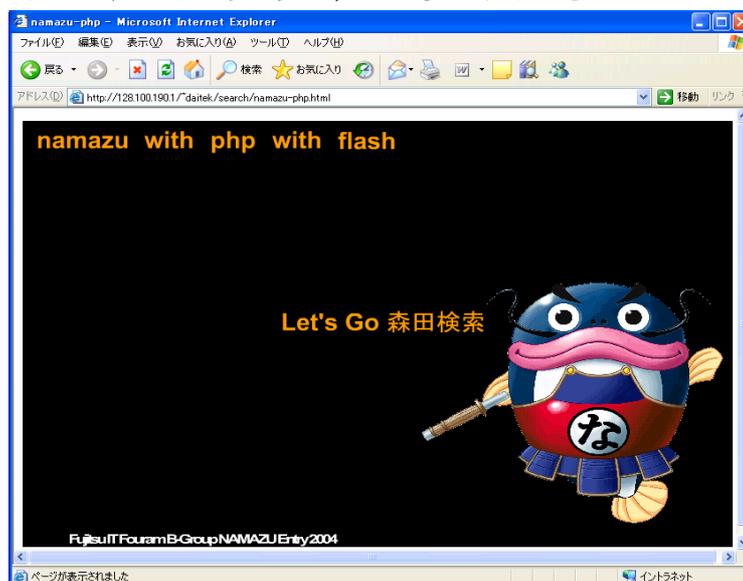


図 1. 『森田検索』操作画面<トップ画面>イメージ図

3. 1. 1. 2 AND 検索イメージ

画面に検索範囲のイメージと検索項目を二つ設け、項目の間に検索条件である AND, OR, NOT のセレクターを配置した。入力が完了したら画面右下の「GO Search」をクリックすると検索結果画面に遷移する仕掛けとした。検索項目の数や検索条件の内容、組み合わせを増やすことで更に複雑で高度な検索が可能となる。

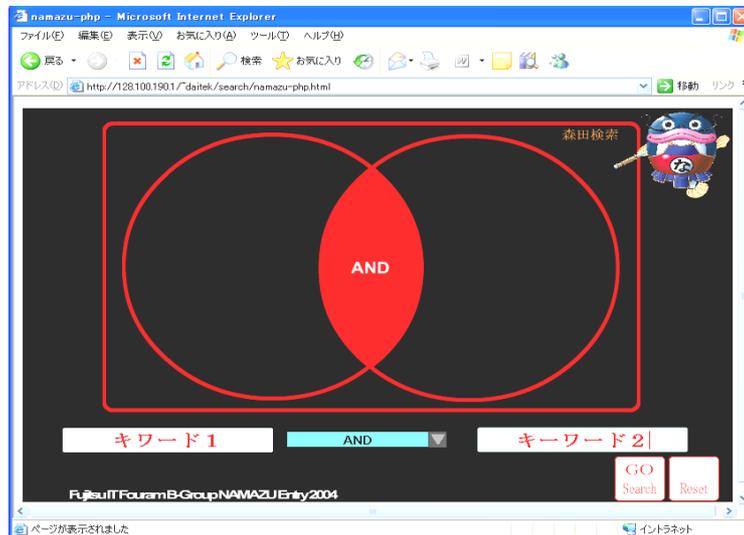


図 2. 『森田検索』操作画面<AND 検索>イメージ図

3. 1. 1. 3 その他 OR, NOT 検索イメージ

以下は、今回開発した OR, NOT 検索のイメージである。それぞれに検索範囲は視覚的に把握できるようになっている。

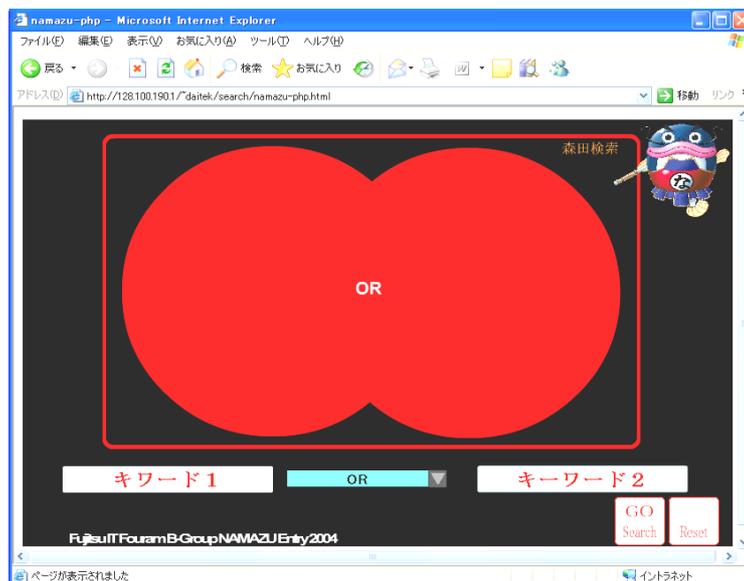


図 3. 『森田検索』操作画面<OR 検索>イメージ図

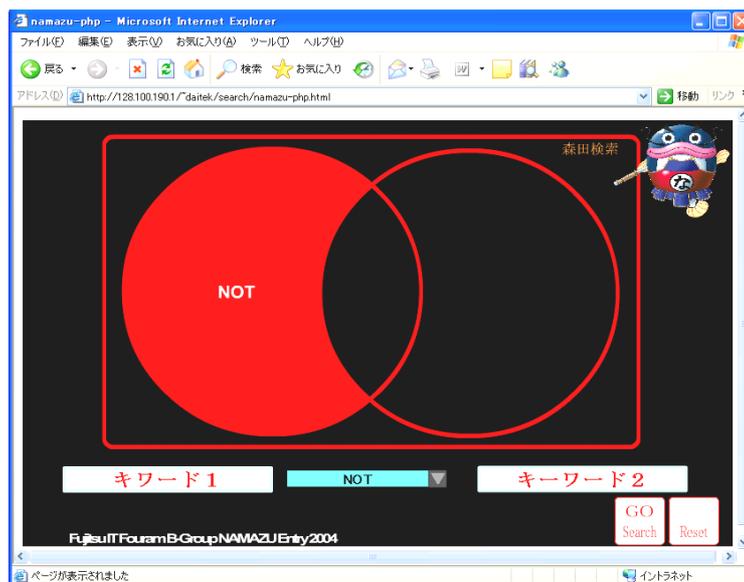


図 4. 『森田検索』操作画面<NOT 検索>イメージ図

3. 1. 2 誰にでも管理できるシステム

次に、Namazu の検索に重要となる、検索結果を表示するために必要なインデックスデータの更新については、問題点でも述べたように知識と手間が必要とされる。

このことを解決するために運用を自動化するシステムの開発を行うこととなった。

- (1) ターゲットのアドレス及びディレクトリの管理
- (2) Namazu のコマンドの自動生成
- (3) 自動収集の実行

このシステムにより、Namazu の知識を必要とせず、だれでも簡単に負担無く管理を行うことができる。詳しくは後に述べるが、私達はこのシステムを『wget (ダブルゲット) パラメータ設定システム』と名づけた。

また、このシステムの実行指示をクローン (Linux の自動実行機能) に組み込むことで、手放しの自動運用が実現できるようにした。

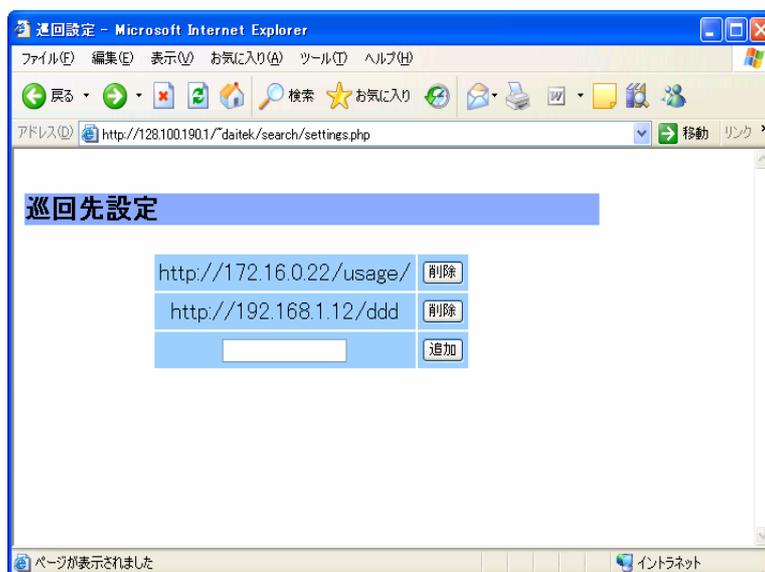


図 5. 『wget パラメータ管理システム』画面図

3. 1. 3 開発における考慮点

このシステムの開発方針が決定されたのは、11回開催されるフォーラムのうちの第5回に決定した。発表前約3ヵ月である。定例の会合は残り5回でこの間に環境の設定、システムの完成と最後の研究成果発表の資料、発表の段取りまでのすべてを終わらせる必要があった。

また、開発に費やせる時間も日常の業務の合間で個々に進めたが時間調整は厳しく、各メンバー非常に苦慮した。全体のすり合わせは臨時の会合を含め（成果発表会を除く）計8回の会合で最終まで仕上げる事が出来た。

スケジュールは以下のようなものである。

- 第5回 9月 9日 方針決定
- 第6回 9月21日 PHP と Namazu 連携確認、PHP 拡張モジュール追加確認
- 第7回 10月 5日 開発担当分担決定
- 第8回 10月19日 開発時範囲決定、開発詳細決定
- 第9回 11月 9日 検索条件入力 (FLASH 開発) , FLASH・PHP・Namazu 連携テスト
- 第10回 11月25日 検索条件入力 (FLASH 開発) , WGET 管理システム開発, テスト
成果発表会に備えて臨時打ち合わせ2回

このように、厳しい時間配分の中仕上げる事が出来た要因には、開発前にグループ内で持っているスキルを把握し、言語の選択、開発範囲を限定したことがポイントとなった。

3. 1. 3. 1 言語の選択と開発範囲の限定

まず、最初のポイントは、Namazu 本体は全くカスタマイズしなかった事である。Namazu のプログラム言語は Perl であり、一般的に熟練を必要とする言語として認識されている。また、Namazu は画面編集と検索プログラムが一つの流れの中に組み込まれている為、一度ミスをおかして動作がおかしくなれば、大きなロスを引き起こすと予想した。更には、私達のメンバーには Perl の熟練者がいないという経験者不足の面からも、リスクを避け、PHP という HTML への組み込み言語を活用することにした。この言語は初心者にも分かり易く、強力な関数と短いプログラミングで済む点も大きなポイントとなった。

関数には Namazu 連携のためのものが提供されており、それらのモジュールを追加して開発することにより短いステップで Namazu の検索結果画面を作成することが可能となる。

よって、Namazu を全くカスタマイズする必要が無く、Namazu の役割としては PHP からのリクエストに回答を返すのみの位置づけとした。

メンバーにも PHP 経験者が複数いたことから PHP が有効であると考えた。

開発時間が厳しい状況で有ったため、リスクをはじめからできるだけ回避して開発に当たったことが短期間で開発できた最大の要因である。

4. システムについて

4. 1 システム構成

- (1) wget パラメータ設定システム
PHP で開発
- (2) 検索エントリー画面
FLASH で開発
- (3) 検索結果表示画面
PHP で開発
- (4) Namazu
ノンカスタマイズ

4. 2 プログラム間連携フロー

プログラム間連携フローは、下記の図6に示す

- (1) 検索エントリーにキーワード入力し検索実行する.
- (2) キーワードを PHP で認識できるコードに変換して渡す.
- (3) 検索結果表示画面は受け取ったキーワードを、一旦、Namazu に検索要求する.
- (4) Namazu は検索要求を受け、インデックスデータから検索し、検索結果データを検索結果表示画面へ返す.
- (5) 再び検索結果画面が Namazu より受け取ったデータを元に結果表示画面を構成しディスプレイする.

このようにしてデータを連携し、結果を得る仕組みとしている. 通常は Namazu だけで検索し検索結果を得られるが、コンセプトを重視することでこのような構成に落ち着いた.

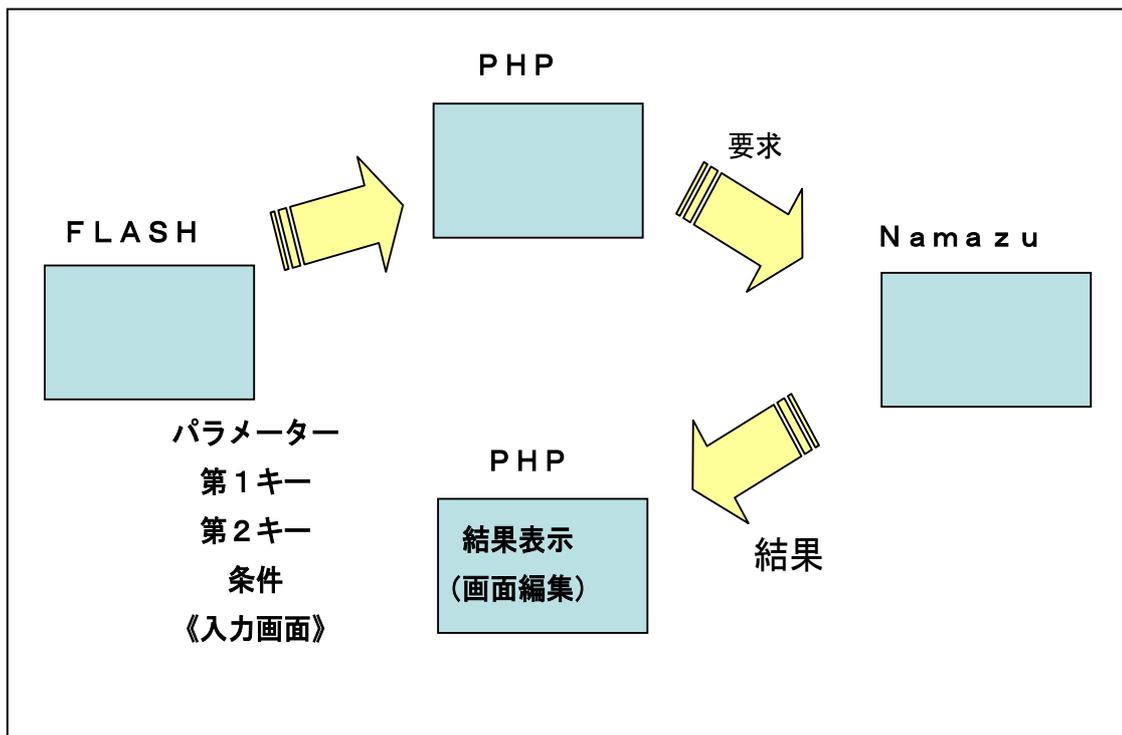


図6. プログラム間連携フロー図

4. 3 wget パラメータ設定システムによる情報収集イメージ

Wget の特徴は、下記の図 7 に示す

このシステムは、目標の URL を簡易な入力画面から登録しておき、登録された URL データを元に各サーバから wget コマンドでコンテンツ情報を巡回収集し、収集したコンテンツ情報から Namazu の検索インデックスを自動生成、生成後は不要となったデータをクリアして完了するまでの完全自動化システムである。

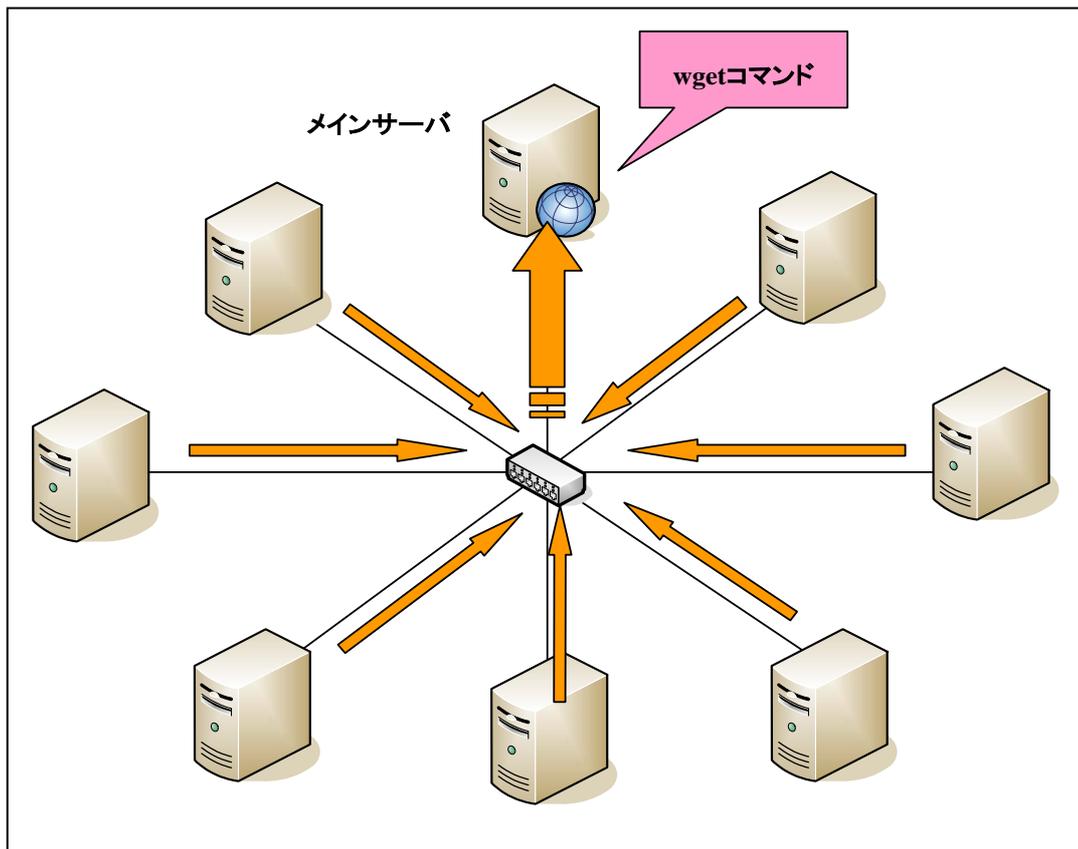


図 7. wget パラメータ設定システムによる情報収集イメージ図

4. 4 INDEX 作成のイメージ

INDEX 作成のイメージは、下記の図 8 に示す

wget で収集したコンテンツ情報を元に Namazu の Index データを作成する。
この作業完了後、wget したコンテンツデータを破棄する。

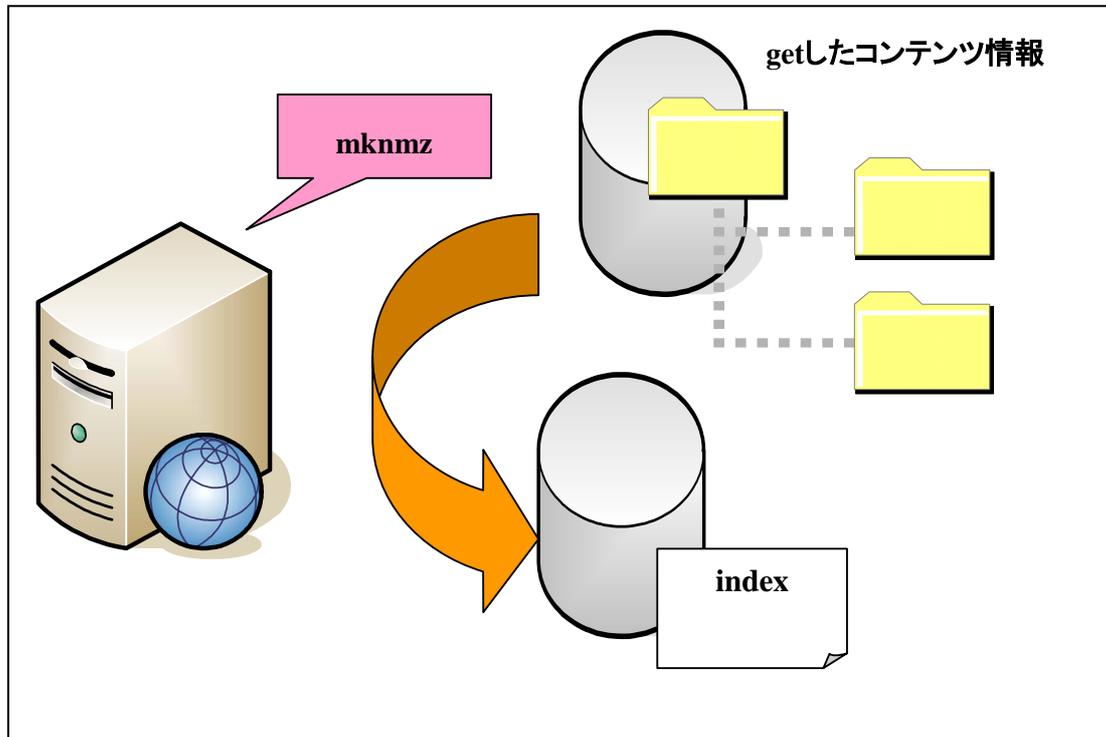


図 8. INDEX 作成のイメージ図

4. 5 利用者が検索したときのイメージ

利用者が検索したときのイメージは、下記の図9に示す

エントリー画面からキーワードを入力し検索を開始する。キーワードは、一旦検索結果画面に渡され、更に Namazu へと引き継がれる。Namazu は検索を行った後、結果データを検索結果表示画面へ返し結果表示をさせる。利用者は検索結果画面で表示された内容を参照し、リンク先のコンテンツ情報を参照できる。

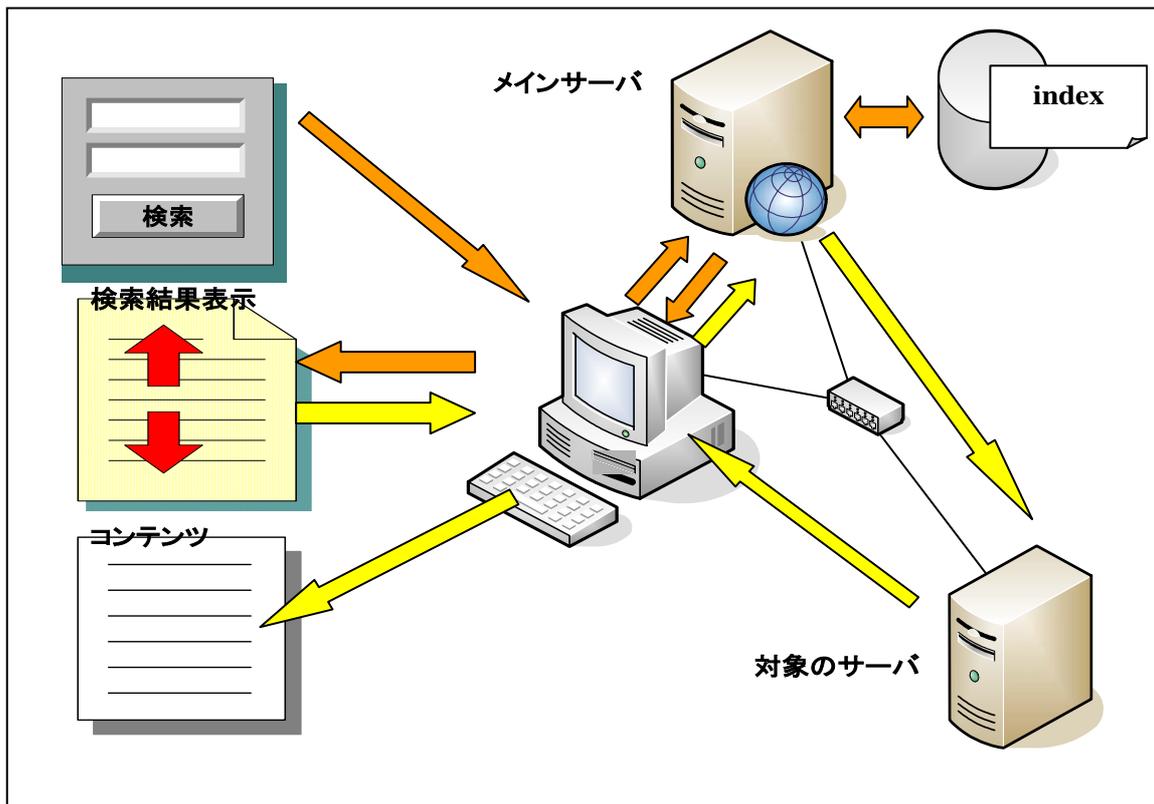


図9. 利用者が検索したときのイメージ図

5. 開発における問題点

開発においては簡単に結果を得られたわけではない。幾つもの問題をクリアして答えを得ながら開発を進めた。以下にトラブルを起こした事象について記載する。

5. 1 Namazu 検索結果の文字化け

この現象は、Namazu を Windows クライアントで表示すると文字化けしてしまう現象が起きた。原因としては Apache の言語対応に必要な設定が漏れていたことで引き起こしていた。扱う文字コードの違いによりこのような現象が引き起こされてしまうのだ。

5. 2 検索結果がすべて 0 件で返される

この現象は、確実に存在するキーワードで、検索エントリー画面から検索を実行しても検索結果を得ることができない現象が起きた。

この原因は、FLASH と PHP 間の文字コードに関係があった。

FLASH バージョン 6 以降は UTF-8 (バージョン 5 以前は SJIS) を扱う仕様に変更され、PHP は EUC-JP を扱う環境にある。

この異なる文字コード間で全角文字をキーワードで扱うには、お互いに文字コード変換してデータを渡さなければならない。

この時点で今回は FLASH MX を使用しており UTF-8 から EUC-JP へコード変換する必要があったが、私達は、UTF-8 ではなく SJIS から EUC-JP へ変換をしていたことが問題であった。

5. 3 見えないエラー

今回のシステムでは PHP の拡張モジュールを追加し、Namazu との連携を実現している。この環境を作る際に PC 単位でインストールできるものできないものが発生した。大まかな原因としては、今回持ち寄った PC はハードディスク容量が十分でないものもあり、Linux をフルインストールしていない環境もあった。このような環境下では今回必要であるデータがインストール済みになっていないことが判明した。しかし、それは技術アドバイザーの手を借り経験値で原因を導き出すことができた。Windows 環境に慣らされた私達だけに思いもよらない現象に改めて驚きを感じた。このような、様々なオープンソースを組み合わせるシステム化するには、エラーの切り分けが、非常に苦勞する点であると実感させられた。

6. システムの評価

システム全体の評価としては、短期間での開発ではあったが、オープンソースとそれに対するアイデアがうまく織り交ざった満足のいく結果を得られた。実用にも十分耐えうる内容で、コアとして今後応用するのによい発想のシステムが出来上がったと評価している。使用した感覚も、非常に違和感なく目標であった親しみのある誰にでも仕えるシステムというイメージにぴったりのシステムに仕上がったことは、当初を振り返ると想像もつかない成果である。

7. 今後の課題

今回のシステムは開発期間の問題もあり、検索キーワード 2 個、検索条件 (AND, OR, NOT) の 3 通りのみの対応になっているが、更に複雑な検索を実現するためには、キーワード指定個数の追加、検索条件を更に高度なものを選択できるようにするなど、今後、更に業務で役立つシステムに発展させることができると期待している。

また、各社各様ではあるが実運用への提案も今後の課題であろう。

8. おわりに

今回の平成 16 年度『IT 研究フォーラム』に集まったメンバーは、営業職～技術職まで、幅広い立場や分野から集まり、意見を出し合い、今までできなかった面白い発想でシステム開発できたことは、もう一つの大きな収穫であった。それぞれのノウハウや環境の異なる職場から見たシステムへの思いが集約して面白い結果を生んだことも、今回の開発の重要

点であるだろう。今後のオープンソースの流れはますます実業務に恩恵をもたらしてくれるものと考えている。

しかし、反面、今回得た教訓の一つは、オープンソースであるからこそ組み合わせには十分に注意が必要という点だ。安定稼動と短期開発を実現するためには、それらに関する情報収集に努力を惜しまず、ノウハウの蓄積に努めることが重要であると再認識することができた。今後、益々オープンソースの発展に目が離せなくなった。

最後に、開発した『森田検索』『w g e tパラメータ管理システム』の画面と森田検索のマスコットを紹介してこの論文を締めたいと思う。

参考文献

- [1] 10日でおぼえるRed Hat Linux 9サーバ構築・管理入門教室
松本光春 著，翔泳社
- [2] Namazuシステムの構築と活用～日本語全文検索徹底ガイド～
馬場 肇 著，ソフトバンク・パブリッシング社