3　本調査

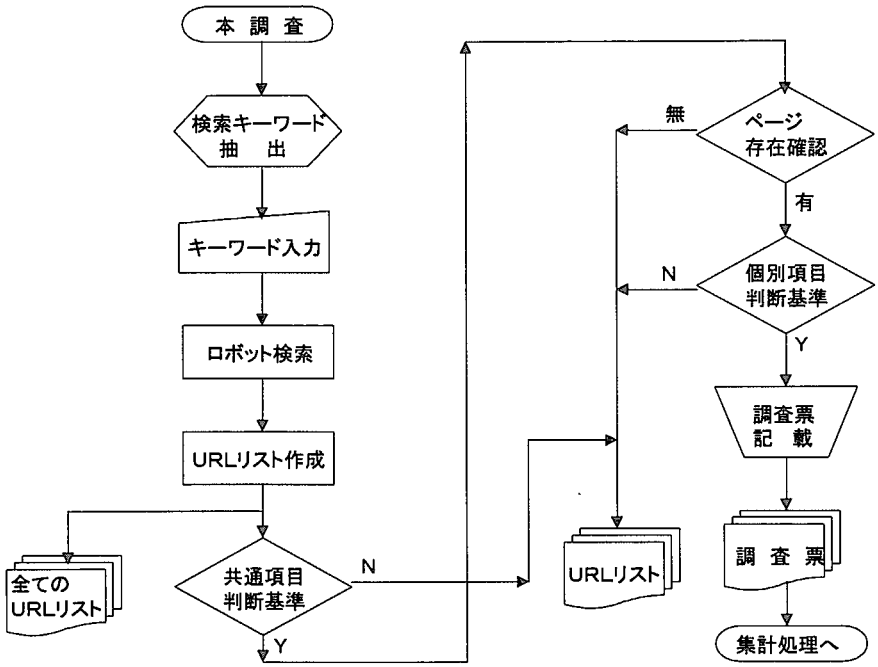

図８　本調査フロー

(1) 検索キーワードの設定

インターネット上のコンテンツの中から違法・有害情報を含むキーワードを抽出する。

キーワードは、予備調査の検索結果を踏まえた上、検索後に該当したＷｅｂの中から適当かどうかを実際に検索を繰り返しながら決定した。

また、実際にＷｅｂ上に記載されていた参考フレーズの中から、キーワードやフレーズとして使えそうなものを抽出して、検索のキーワードとした。

＜参考フレーズの一例＞

① 元手なしのボロ儲けＢ勘定屋の儲けのカラクリを徹底調査
② 合法的に飲み屋のツケを踏み倒すノウハウを公開
③ スピード取締りカメラ防御用マジックナンバープレートの製作方法を検証
④ 悪用厳禁タダで郵便ものを出す悪の手口を公開
⑤ 大手ＤＭ会社も絶賛ダイレクトメールを絶対に開封させるテクニック
⑥ 盗聴器がなくても隣の会話を簡単に盗み聴くことができる盗聴術を伝授
⑦ あなたの側にもいる覚醒剤、麻薬中毒患者を一目で見破る方法を伝授
⑧ 水不足解消合法的に水道料金を半分にする裏テクニックを公開
⑨ これで発音はバッチリあなたの言葉を伝え易くする簡易発声練習表を公開
⑩ ニセ札をつかまされるな警察も使っている秘密兵器を紹介

(2) ロボット検索

今回の「インターネット実態調査」は、調査時点におけるインターネットの状況を知ることが目的である。このため、ロボット型のサーチエンジンの一つである「UltraseekServer」を利用し、独自に検索のためのシステムをインターネット上に構築した上、８つのカテゴリー毎に用意した検索キーワードにて、サーチエンジンを稼動させ当該ＵＲＬの収集を、実施した。

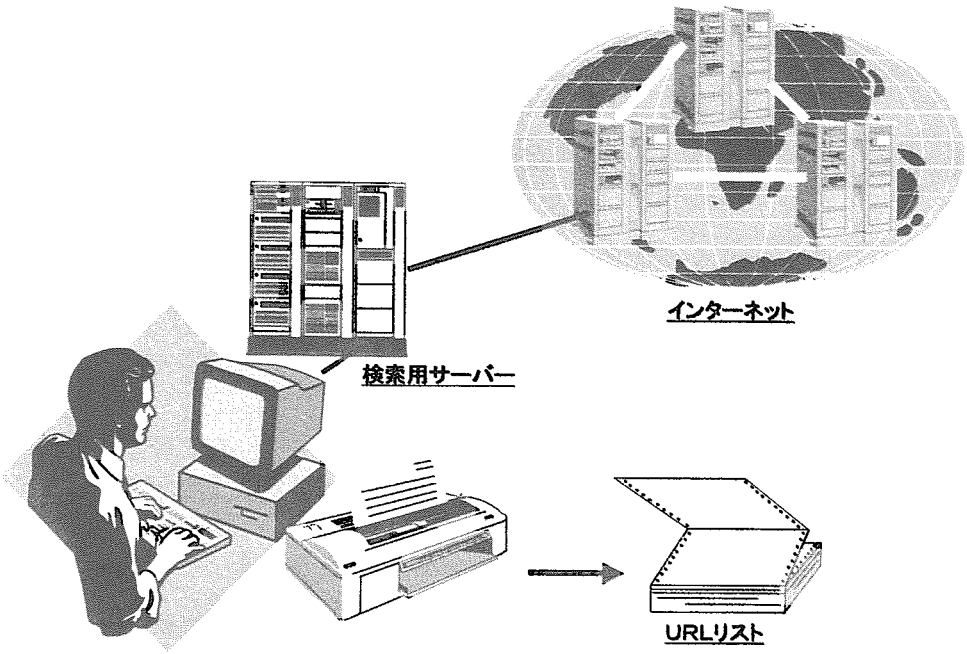

図９　本調査の構成

なお、収集対象サイトは、日本語サイトを中心に行うため、ｊｐドメインを対象としたサーチエンジンの設定を行ったが、リンク等があった場合はｃｏｍドメインも対象とした。

さらに、ロボット型サーチエンジンを利用した検索における留意点としては、①ｒｏｂｏｔｓ．ｔｘｔ（ロボット・テキスト）の参照、②ＭＥＴＡｔａｇ、③当該サイトによる対抗手段（逆アタック）などがある。以下、それぞれ簡単に説明する。

① ｒｏｂｏｔｓ．ｔｘｔ（ロボット・テキスト）

ロボット型サーチエンジンの動作を規定するために、それぞれのＷｅｂサイトが自サイトに用意するファイルである。検索ロボットはこのファイルに書かれている内容にしたがって、そのウェブサイト内のコンテンツを収集してよいかどうかを判断することが求められる。違法・有害情報を提供しようとするも

のがこのファイルにより、ロボットでの検索を無効化していることも十分考えられるため、独自ロボットでの収集においては、このファイルの扱いに十分注意する必要がある。

② ＭＥＴＡｔａｇ

それぞれのページに書き込まれている情報のひとつである。それぞれのサーチエンジンは、このＭＥＴＡｔａｇに書き込まれた情報をファイルの検索に利用している。ｒｏｂｏｔｓ．ｔｘｔと同様、注意して扱う必要がある。

③当該サイトによる対抗手段（逆アタック）

ページを閲覧した記録（ログファイル）から、こちら側が情報収集を目的にファイルを調査していることに対抗して、こちら側の検索ロボットを搭載したサイトへの不正クラックが行われる可能性がある。このため、今回の調査では、ファイアウォールを設置するなどして、セキュリティの確保に注意した。

このように、今回の調査においては、調査自体の秘匿性をある程度保つとともに、一般に使われているロボット型サーチエンジンを超える範囲まで調査対象としている。

ロボット型のサーチエンジンによって検索対象となるＵＲＬは最大数百万 URL もの多数にのぼるため、キーワードによる絞込みを行ったとしても、依然として多数のＵＲＬを検出する。

これらのＵＲＬについては、該当するキーワードを含んでいるだけで、内容が伴わないものも多数含んでいるため、収集したＵＲＬの確認作業は、カテゴリー毎の検索得点による確認作業基準点を設け、この基準を満たしたものをを確認する他、収集した全ＵＲＬに対して、タイトルを目視チェックし、特に不信な内容を含んでいる可能性をもつＵＲＬについても確認することとした。

検索得点による確認作業基準点を設けたのは、確認作業の際に、独自サーチエンジンがリスト化したデータから、キーワードへの適合度の大きさが、本調査で確認すべき内容を含んでいるかどうかの判断の目安になるためである。

ただし、カテゴリー毎の検索キーワードにより、サーチエンジンで得られたキーワードに適合するＵＲＬの件数にばらつきが起こるため、ヒット率の悪いカテゴリについても、ある程度、まとまった数量のＵＲＬについて確認を行えるよう、基準点を設定した。

なお、検索得点による基準の詳細は、以下のとおり。

①薬物関連	98％まで
②銃器関連	96％まで
③悪質商法関連	98％まで
④賭博関連	98％まで
⑤著作権関連	98％まで
⑥性風俗関連	98％まで
⑦名誉毀損関連	97％まで
⑧犯罪誘発関連	97％まで

＜UltraseekServerの検索得点の仕組みについて＞

検索得点の計算に入る項目は、

①ドキュメント中のキーワード使用頻度数

②ドキュメントの長さ

③キーワードの場所

④フレーズの正確性

である。

例えば、キーワードが、長いドキュメントに５回ヒットよりも、短いドキュメントに４回ヒットした方がスコアが高くなる。

また、タイトルやサマリー、ｂｏｄｙタグ中のテキスト等キーワードの場所によっても検索得点の倍率が違い、検索得点に反映されている。

さらに、フレーズで検索を行った場合、フレーズの正確性（フレーズを、単語に分解後、一つの単語が何回も出るよりも、フレーズが全部ヒットした方が高いスコアになる。）も反映されている。

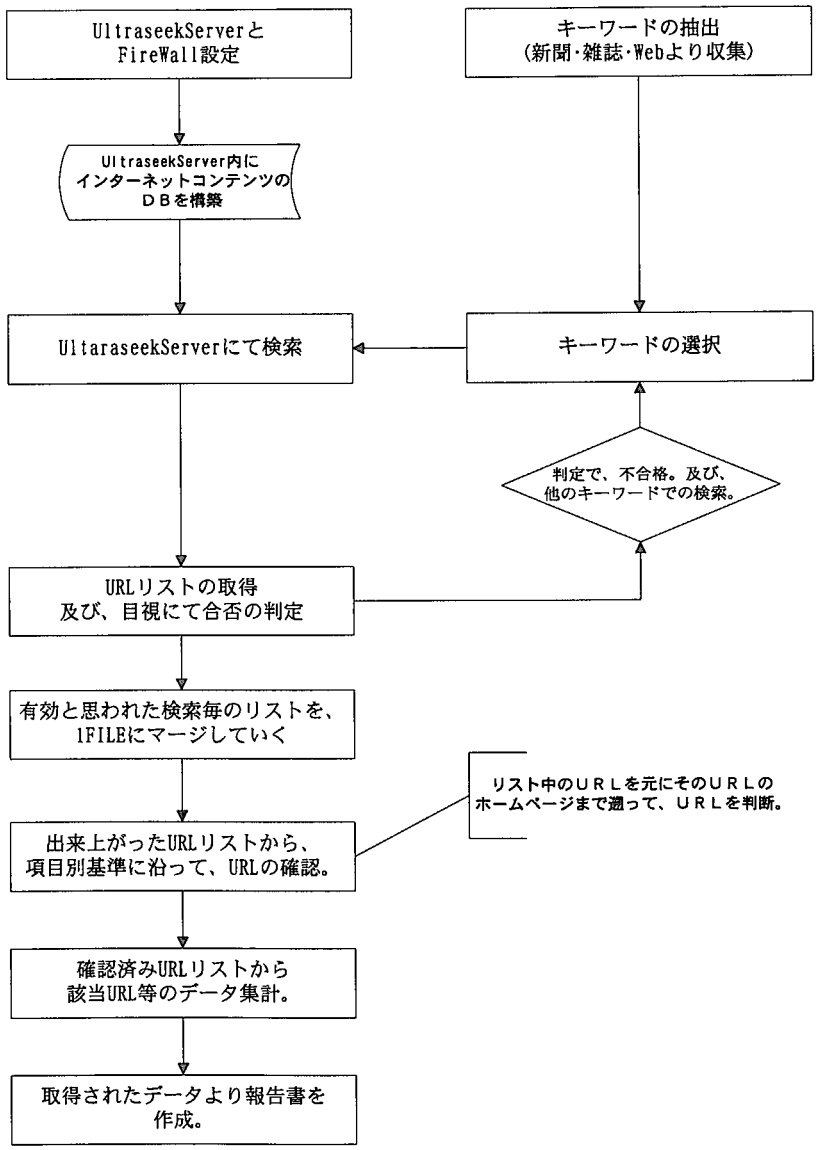

図１０　作業フローチャート

(3) ＵＲＬリスト作成

ロボット検索で該当した全てのＵＲＬリストを作成する。

(4) 共通項目判断基準

該当したＵＲＬリスト全てを目視で確認することは、非常に効率が悪いので、当該ＵＲＬリストの中、共通項目判断基準（検索得点等：第３＿４項（１）共通項目参照）によりフィリタリングを行う。

(5) ページの存在を確認

ロボット検索により該当したＵＲＬでも、確認作業時には、そのコンテンツが消去されている場合があり、これらは除外する。

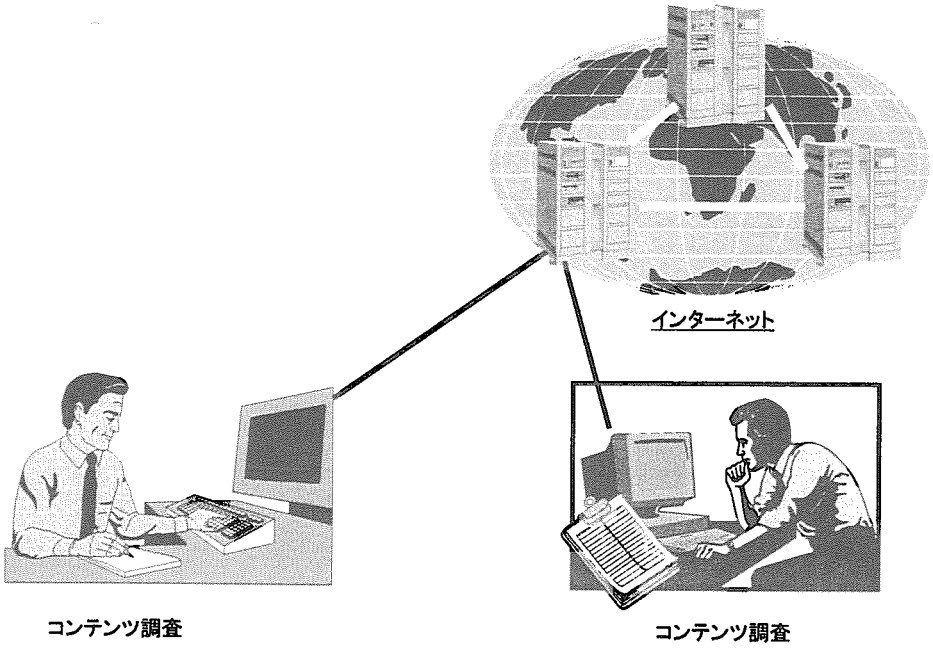

図１１　本調査作業状況

(6) 個別項目判断基準

前項により、抽出されたＵＲＬにより目視でコンテンツの内容を確認し、各カテゴリー毎の判断基準により、関連サイトに該当するか否かを決める。

各カテゴリー毎の判断基準は、第３＿４項　調査事項（２）個別項目で列記している。

(7) 調査票記載

コンテンツの内容等を確認し、調査票の調査項目に従い、記載する。

各カテゴリー毎の調査項目は、第３＿４項　調査事項（２）個別項目で列記している。

使用した同調査票の様式は、第６＿２項を参照のこと。

(8) 調査結果

ア　ＵＲＬリストにより、各コンテンツの内容を確認し、チェック票に基づき分類、統計化を行う。

イ　チェック票を集計し、インターネット上の実態を定量的に分析する。

ウ　調査結果に対する検討及び提言等を加え報告書を作成する。