別紙

データロード実験結果

1　データロード作業

1．1　作業の概要

平成 25 年 12 月 12 日から 12 月 25 日まで、国立情報学研究所内の PC を使用して、全件データロード作業を行った。対象としたデータは合計 29,369 件である。データは WEKO の一括登録ツールである SWORD Client for WEKO (SCfW)を使用して、2,000 件ずつ一括でアップロードを行った。

データロードは合計で 53 回実施し、うち 38 回でエラーが発生した。リトライはデータ容量の大きな本文データのロードにて発生した。そこで、当該ファイルのロードが成功するように、サーバー側の Apache、PHP、MySQL の設定ファイルのパラメータチューニングを実施したところ、エラーの発生回数が減少した。データロードの平均登録時間は、2,000 件あたり約 68 分であった。

1．2　全件データロード処理

全件データロードは29,369件を10ロットに分割して行った。データロード記録は下表のとおり。

ロット	件数(件)	合計処理時間(分)	リトライ回数	リトライへの対処
1	2,000	97	11 回	サーバーのチューニングを実施
2	2,000	56	0 回	－
3	2,000	41	0 回	－
4	2,000	43	0 回	－
5	2,000	89	0 回	－
6	2,000	75	2 回	大容量ファイルが主な要因。問題となった大容量（383MB）のファイルを一括登録によらず Web 入力画面から登録
7	2,000	66	0 回	－
8	2,000	48	0 回	－
9	2,000	51	0 回	－
10	2,000	152	19 回	サーバーのチューニングを実施
11	2,000	58	3 回	サーバーのチューニングを実施
12	2,000	51	0 回	－
13	2,000	63	0 回	－
14	2,000	63	0 回	－
15	1,369	49	3 回	大容量ファイルが主な要因。データの一括登録数を 500 件ずつに減らして実施
合計	29,369	1,002	38 回	－

２　データロード結果と不具合への対処

２．１　結果概要

データロードの結果、メタデータの不具合が 7 種見つかったが、いずれも対処方法は検討済である。なお、メタデータの不具合の詳細は２．２で述べる。

また、メタデータの不具合以外にも、いくつかの不具合や不都合が発生した。内容は以下の通りである。

No	不具合内容	原因	対策
1	データロード中、MySQL の処理が終了しない	大容量のファイルの登録	サーバーのチューニング
2	データロード中、HTTP のタイムアウトが発生することがある	大容量のファイルを登録する際に時間がかかる	一括登録によらず Web 入力画面から登録する
3	インデックスツリーの表示が遅くなる	インデックスツリー数が多い	対策中
4	インデックスツリーのパスの長さに制約がある	Windows のシステム上の制約	コンバータを改修し、長いインデックスツリーをチェックしアラートを出すよう改修

２．２　メタデータに関する不具合とその原因、対処方法一覧

移行実験で発見したメタデータの不具合については、以下のとおり対策案の検討を行っている。

No	不具合内容	原因	対策
1	刊行年月日の月日が欠落	不正な日付データ	元データが不正なため、対処しない
2	入力データの時刻が欠落	刊行年月日（日付型）に時刻を入れようとした	刊行年月日に時刻は不要なので、対処しない
3	フィードバックメールのアドレスがない	フィードバックメール管理機能が未実装の状態でメタデータを編集した	フィードバックメール管理機能は実装済であり、今後は発生しない
4	本文ファイルが登録されていない	ファイル名の文字コードの変換失敗	SCfW でアイテムを登録する際、UTF-8 指定で登録を行う
5	抄録が分割登録されている	抄録に、区切り記号”|”が含まれていた	データに”|”が含まれていないか、コンバータにチェックする機能をつけることを検討
6	データの重複登録	オペレーションミス	-
7	抄録に含まれている”C”の文字以降が登録されていない	”C”が MySQL で登録できない 4 バイトの UTF-8 の文字だった	データに 4 バイトの UTF-8 の文字が含まれていないか、チェックする機能をつけることを検討