УДК 81'322.2

Анализ тематических кластеров текстовых коллекций и исследование временно`й динамики тем (на материале конференций по Argument Mining)

Пименов И.С. (Новосибирский государственный университет),
Саломатина Н.В. (Институт математики им. С.Л. Соболева),
Сидорова Е.А. (Институт систем информатики им. А.П. Ершова)

В статье представлены результаты исследования изменений, происходящих в тематических кластерах, построенных на коллекции текстов конференций предметной области Argument mining. Выявление терминов, установление связей между ними и тематическая кластеризация проведены с помощью сторонних программных средств, позволяющей извлекать термины в форме именных словосочетаний, проводить их кластеризацию на базе алгоритма, основанного на применении функции модулярности. Приводится оценка качества полученных кластеров по трем критериям. Трансформацию терминологического состава кластеров во времени предлагается анализировать с помощью ориентированных графов, построенных на основе критерия, который позволяет фиксировать наиболее важные изменения. Терминологическая лексика выявленных тематических кластеров характеризует отдельные направления, в которых ведутся исследования, а трансформация терминологического состава кластеров во времени демонстрирует смещение интересов.

Ключевые слова: *тематическая кластеризация, коллекции текстов предметной области, динамика тем во времени, Argument Mining*

1. Введение

Исследование временно`й динамики тем различных предметных областей (ПО) представляет интерес не только для решения масштабных задач: отслеживания эволюции ПО, построения прогнозов дальнейшего развития, но также и утилитарных, таких как знакомство с ПО, обновление профессиональных знаний и пр. Данная работа направлена на создание инструмента для поддержки исследований такого типа, например, в помощь при построении реферативных обзоров для того, чтобы выявить актуальные направления ПО, применяемые в них методы. Автоматическая обработка коллекций текстов дает возможность

увеличить объем анализируемой литературы и позволяет ослабить субъективное влияние экспертов ПО.

Коллекции текстов, как правило, формируются из статей, полученных по запросам к базам данных сетей цитирования и\или из рейтинговых публикаций в тематических журналах. Особенность данного исследования заключается, в частности, в том, что оно проводится на материале трудов конференций. Для них, в отличие от журнальных статей, как правило, рейтинг неизвестен. Но статус конференции и сам факт прохождения отбора говорит о значимости анализируемых текстов. Следует признать, что риск ошибок в оценивании тематических изменений ПО по материалам конференций выше, чем по рейтинговым статьям. Тем не менее, такое исследование представляет интерес, поскольку позволяет оперативно получать представление о смещении научных интересов на самом раннем этапе.

Структура ПО, задаваемая тематическими кластерами, может быть построена путем применения к текстовым коллекциям любых известных методов кластеризации (иерархических, k-средних, плотностных, графовых и пр., см., например, [1]), используемых для анализа нетекстовых типов данных.

Подходы к кластеризации можно разделить на кластеризацию текстов по *совместной встречаемости* в них *терминов* (co-word analysis), например, [2] и по *совместному цитированию* (co-citation analysis), в частности, упоминанию в одной публикации пары двух других или, наоборот, цитированию одной и той же публикации другими работами [3, 4]. Как показано в работе [5], результаты структурирования ПО, полученные с применением этих двух подходов к одним и тем же данным, близки.

Исследование динамики изменений, происходящих в тематических кластерах с течением времени, проводится, как правило, с установлением всех возможных изменений в составе терминов. К примеру, в работе [6] применяется метод скользящего окна, структура кластеров фиксируется в перекрывающихся временных интервалах, что позволяет максимально подробно фиксировать зависимость состава кластеров от большого числа параметров. В других работах, таких как [5], применяется такой показатель как индекс включения, не дифференцирующий типы происходящих изменений. В данной работе предлагается критерий для отслеживания только существенных изменений в развитии тематической структуры ПО.

Цели работы: 1) разработать pipeline, позволяющий проводить анализ динамики тематических кластеров во времени; 2) провести апробацию на коллекции текстов ПО.

Результаты, представленные в работе, являются предварительными, поскольку они будут уточняться на коллекции, пополненной текстами исследуемой ПО данного и других временных периодов.

Исследование выполнено при поддержке РФФИ в рамках проектов № 18-00-01376 (18-00-00889) и № 18-00-01376 (18-00-00760).

2. Методы

Методика проводимого исследования включает следующие шаги:

1. Сбор коллекции.
2. Предобработка коллекции (токенизация, морфологический анализ).
3. Извлечение терминов.
4. Структурирование ПО (тематическая кластеризация текстов коллекции).
5. Оценка качества построенных кластеров.
6. Построение графов трансформации тем.

Для относительно новых ПО сбор коллекций является непростым делом, большая часть актуальных текстов часто оказывается вне зоны свободного доступа. Основными требованиями к коллекции текстов для анализа тематической трансформации во времени является примерно равное количество текстов для каждого рассматриваемого периода и жанровая однородность (статья, доклад, тезисы, …).

В качестве инструмента для решения задач, указанных в пп. 2, 3 и 4, для отдельного временного среза применяются методы, которые реализованы в свободно распространяемой программе VOSviewer [7]. Они позволяют проводить кластеризацию терминов как по совместному цитированию, так и по совместной встречаемости терминов (co-occurrence links) в полных текстах. В результате реализуется четкая классификация по терминам ПО и нечеткая по текстам.

Изменения кластеров во времени отражаются в двудольном графе, построенном на основе критерия, позволяющего отслеживать трансформацию тем избирательно.

2.1. Методы, реализованные в программе VOSviewer

Извлечение терминоподобных словосочетаний (в дальнейшем для краткости «терминов») производится по шаблону, определяемому как последовательность существительных и прилагательных, оканчивающаяся на существительное.

Структура ПО формируется в виде сети путем связывания терминов согласно вычисленной силе связи (s_{ij}) между терминами i и j, определяемой мерой ассоциации:

$s_{ij} = 2mc_{ij}/c_ic_j$, где c_{ij} – число текстов, в которых термины i и j встречаются совместно, а c_i, c_j – число текстов, в которых встречается i-ый и j-ый термин соответственно, m – общее число связей в сети [8].

Унифицированный (единый) подход к кластеризации и визуализации, реализуемый программой VOSviewer, опирается на нахождение экстремума функции модулярности, в которой имеется параметр (r), позволяющий кластеризацию сети с необходимой степенью детализации. Термины, объединенные в кластеры, характеризуют отдельные темы ПО. Программа дает возможность отфильтровывать служебную и общеупотребительную лексику по предлагаемому пользователем Стоп-словарю, объединять кластеры с малым количеством элементов с ближайшими более крупными.

2.2. Методы оценки качества кластеров

Известно, что проблема проверки адекватности кластеризации не решена в теоретическом плане. Без априорного знания принадлежности объектов к построенным кластерам задача оценки качества чаще всего решается вручную. Формальные критерии проверки сводятся, в основном, к проверке таких свойств как компактность, концентрация и отделимость [9]. Построенные критерии обычно объединяют проверку свойств компактности и отделимости, проверка концентрации элементов кластера вокруг его центра представлена чаще неявно.

В данном исследовании оценка качества кластеров позволяет подобрать наилучшим образом параметры программы VOSviewer для проведения кластеризации. Были использованы три критерия, по которым производилось оценивание компактности и отделимости строящихся кластеров:

1. Validity index (*CS*) [10] измеряет отношение максимального суммарного расстояния между элементами одного кластера к минимальному суммарному расстоянию между центрами кластеров. Чем меньше значение критерия, тем более качественным считается разбиение на кластеры.

2. Calinski-Harabasz index (*VCR*) [11] вычисляет нормированное отношение суммарного внутрикластерного расстояния к суммарному расстоянию в кластеризуемом множестве. Для "идеальной" кластеризации оно равно 1, т.е. чем ближе значение индекса к 1, тем лучше разбиение множества на кластеры.

3. Silhouette индекс (*SWC*) [12] фактически является мерой несхожести элементов одного кластера с элементами других кластеров и гарантирует лучшее разбиение при высоком значении *SWC*.

2.3. Построение графов трансформации тем

В данной работе для отслеживания временных изменений в кластере используются ориентированные графы, отражающие изменение терминологического состава кластера при переходе из отсчета времени t в t+1.

Пусть $C_t = \{c_i^t\}$ – множество кластеров в коллекции текстов, относящихся к периоду времени t, i = 1, …, n_{cit}, n_{cit} – число терминов в кластере c_i^t.

Пусть при переходе от временного среза t к t+1 каждый кластер c_i^t трансформируется в упорядоченное (по убыванию числа содержащихся в нем терминов отсчета t) множество $\{c_j^{t+1}\}$, j = 1,…, K_j, K_j – число вновь образовавшихся кластеров. В анализе трансформации кластера c_i^t при переходе к отсчету t+1 учитываются четыре типа преобразований, отвечающие следующим условиям:

1) кластер c_i^t трансформировался большей частью в c_1^{t+1}, так что $\Delta_{1,2} > p$ (p – порог, регулирующий объем пересечения кластеров);

2) кластер c_i^t трансформировался большей частью в c_1^{t+1} и c_2^{t+1}, так что $\Delta_{1,2} \le p$ и $\Delta_{1,3} > p$.

3) число кластеров, в которые трансформировался большей частью кластер c_i^t, превышает два (j > 2 и $\Delta_{1,j} > p$);

4) кластер c_i^t отсутствует во множестве $\{c_j^{t+1}\}$ большей частью своих элементов: $\Delta_{0,1} > p$ ($c_0^t = c_i^t \setminus c_j^{t+1}$), где $\Delta_{1,n} = |c_1^{t+1} \cap c_i^t| / |c_i^t| - |c_n^{t+1} \cap c_i^t| / |c_i^t|$, (n > 1).

Тип 1) можно интерпретировать как сохранение темы, возможно, с обновлением, 2) – как выделение самостоятельной темы. Типы 3), 4) – это два типа прекращения существования темы (поглощение другими кластерами, практически полное исчезновение). Выполнение условий обеспечивает отбор кластеров, играющих существенную роль в трансформации.

3. Эксперимент

Обработка коллекции текстов ПО Argument Mining (токенизация, морфологический анализ, извлечение терминов, кластеризация) для данных каждого временного отсчета (2015, 2016, 2017 гг.) выполнена программой VOSviewer. Большая часть служебных и общеупотребительных слов отфильтрована по собранному авторами Стоп-словарю, содержащему более 500 слов и словосочетаний. Оценка качества построенных кластеров позволяет считать проведенную кластеризацию удовлетворительной. Изменение терминологического состава кластеров зафиксировано в графе развития тем, что дает возможность выявлять исследовательские приоритеты для данной ПО и каждого временного среза.

3.1. Используемые данные

В анализируемую коллекцию всего рассматриваемого периода (2015 – 2017 гг.) вошел 51 доклад, сделанный на конференциях WorkShop 2015, 2016, 2017: 16 докладов, 20 докладов и 15 докладов соответственно, общим объемом около 184 тыс. словоупотреблений. Из текстов докладов были удалены ссылки на литературу, параграфы, посвященные обзору литературы.

3.2. Результаты эксперимента

Выявленные программой VOSviewer термины представляют собой, в основном, одно- и двух (реже трех-) словные именные группы с текстовой частотой $f_t > 2$. В первом периоде таких сочетаний оказалось 209, во втором – 308 и 216 – в третьем, всего за весь период 2015 – 2017 рассмотрено 443 разных термина. В результате кластеризации терминов в отдельные периоды получены 4, 5 и 3 непересекающихся кластера соответственно, объемом от 25 (2015 г.) до 97 (2016 г.) терминов.

3.2.1. Содержание тематических кластеров

Терминологический состав каждого кластера характеризует подход к исследованиям в целом: параметры используемых текстовых ресурсов (темы и объемы данных), применяемые методы (включая информацию об анализируемых текстовых фрагментах), распознаваемые объекты (см. таблицу 1). Кластеры в таблице пронумерованы и упорядочены по убыванию объема содержащихся в них терминов.

Таблица 1. Термины, характеризующие кластеры докладов 2015, 2016, 2017 гг.

временной отсчет	№ кластера	термины		
		данные	методы	объекты распознавания
2015 г.	1	opinion, corpus, corpora, wordnet, …	dataset, training set, annotation process, expert, indicator, semantic similarity, latent dirichlet allocation, machine learning, topic model, unigram, tree, …	claim, major claim, argument (-ative) structure, argument component, premise, attack relation, …
	2	collection, document, ...	sentence, phrase, token, word, verb, noun, lemma, vector, weight, model, graph, rule,	negative sentiment, argument, original claim, ...

			pattern, frequency, cosine similarity, score, measure, distance, cluster, ...	
	3	debate, politic, large set, ...	classifier, training data, test set, classification, recall, precision, fold cross validation, n-gram, adjective, adverb, boundary token, iteration, probability distribution, ...	...
	4	article, post	annotation, inter annotator agreement, context, label, argumentation mining, baseline, experimental setup, feature vector, modal verb, ...	argumentation relation, support relation, negation, ...
2016 г.	1	opinion, gay right, debate, online debate, collection, document, article, post, ...	word, phrase, sentence, unigram, model, topic, feature set, sentiment feature, verb, classifier, evaluation, test set, baseline model, context, stance classification, tree, graph,...	sentiment, negative sentiment, ...
	2	persuasive essay, alcohol, ...	annotation process, expert, inter annotator agreement, annotator, expert annotator, annotation guideline, annotation procedure, consensus, proposition, statement, clause, training, occurrence, indicator, rule, distribution, confusion matrix, elementary discourse unit, linguistic feature, ...	argumentation scheme, attack, major claim, support relation, inference, ...

	3	corpus, forum, argumentative microtext, ...	link, node, pattern, frequency, training data, classification, testing, false negative, false positive, recall, precision, f1 score, manual analysis, statistic, cohens kappa, controversial topic, parser, location, segmentation, tree,... structure	Argument (-ative, -ation) structure, argument component, premise, attack relation, negation, rebuttal, ...
	4	data set, obama, marijuana, ...	experimental setup, human annotator, annotation task, token, semantic similarity, score, measure, individual feature, random forest, ...	claim, main claim, ...
	5		title, whole sentence, training set, bigram, n-gram, fold cross validation, svm, support vector machine, binary feature, cross validation, ...	
2017 г.	1	opinion, debate, corpus, report, web, ...	proposition, sentence, noun, human annotator, inter annotator agreement, manual analysis, rule, pattern, distribution, cohens kappa, lda, cosine similarity, distance, character, metric, ranking, semantic, supervised machine, validity, recall, f1 score,...	attack, argument structure, negation,
	2	dataset, collection, corpora, Wikipedia, ...	natural language processing, model, topic, token, word, unigram, classifier, training data, test set, context, stance	claim, sentiment, ...

			classification, prediction, pipeline, baseline, precision, score, svm, probability, logistic regression, tf idf, word embedding, graph, ...	
	3	document, article, argumentative text, persuasive essay, ...	annotator, annotation scheme, expert classification, genre, labeling, clause, feature set, occurrence, confusion matrix, statistic, measure, criterium, error analysis, ...	argument scheme, argument component, argumentative structure, argumentative unit, argumentative relation, major claim, premise, ...

Следует заметить, что в таблице есть ячейки, которые слабо заполнены или не заполнены вовсе. Например, "объекты распознавания" отсутствуют в кластере 3 в 2015 г. и в кластере 5 в 2016 г. Это означает, что термины, относящиеся к этим аспектам текста, встречаются менее, чем в 3-х текстах коллекции и не прошли фильтрацию по текстовой частоте. В кластере 3 это, например, энтимемы, в кластере 5 – противоречивые темы, предложения для аргументации о значении терминов, свидетельства отсутствия оппозиции в эссе.

3.2.2. Оценка качества кластеров

Полученные количественные значения критериев оценки качества кластеров (см. п. 2.2), представленные в табл. 2, показали, что наилучшее разделение терминов на кластеры достигается при значении параметра r (детализация сети терминов) равном 1. При выбранном значении резолюции кластеры с малым числом элементов в нашем случае не образуются и параметр, позволяющий производить объединение мелких кластеров с ближайшими бо`льшими, на результат кластеризации влияния не оказывает.

Таблица 2. Количественные оценки качества точной кластеризации в каждый из временных отсчетов по критериям *VCR*, *CS*, *CWS*

год	резолюция	Число класт.	*VCR*	*CS*	*CWS*
2015	1,00	4	30,708	0,333	0,139
	1,05	5	26,275	1,223	0,040

	1,10	11	12,329	8,204	-0,165
2016	1,00	5	40,109	0,391	0,087
	1,05	6	30,716	0,430	-0,023
	1,10	12	18,464	0,787	-0,090
2017	1,00	3	57,267	0,076	0,313
	1,05	4	40,505	0,597	0,183
	1,10	8	15,761	7,584	-0,007

Оценка качества кластеров при изменении параметра r, регулирующего число кластеров, показала, что при росте r качество кластеризации падает, хотя при r = 1 и r = 1,05 увеличение числа кластеров на единицу не влечет большого изменения состава кластеров, а вызывает лишь отделение одного небольшого по объему кластера. Но уже для r = 1,10 стабильность кластеризации нарушается.

3.2.3. Динамика развития тем

Для построения графа развития тем применяемый критерий (см. п. 2.3) был скорректирован в силу особенностей данных: небольшого объема коллекции и фильтрации терминов по текстовой частоте. Порог p повышен до значения 0,16. В знаменателе критерия учитывалась мощность множества $\underline{c}_i^t = c_i^t \setminus L_i^t$, где L_i^t – множество терминов, отсутствующих во множестве терминов во временном срезе t+1 (вышедших из оборота). Результирующий граф представлен на Рис. 1. Узлы, обозначенные символом N, содержат термины, не встречавшиеся в текстах докладов предыдущего временного среза. Толщина линий пропорциональна объему терминов, общих для смежных временных срезов.

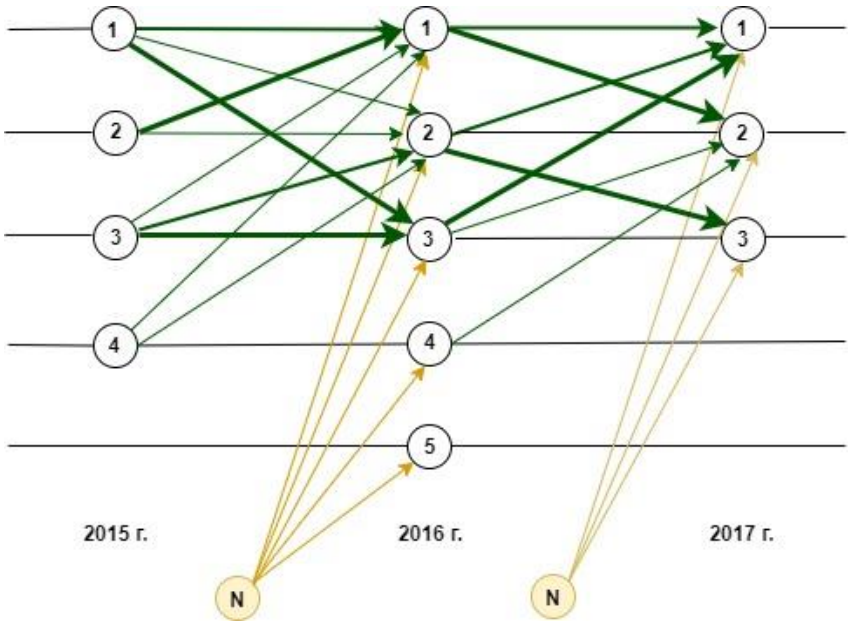

Рис.1. Граф развития тем за период 2015 – 2017.

Причинами нестабильности поведения кластеров, в частности, являются неустоявшаяся терминология ПО, отличающаяся наличием вариантов написания терминов (argument (-ative, -ation) structure; svm, support vector machine и др.), особенности жанра (доклады конференций, где смена участников, а следовательно, и обновление тем, неизбежны). Но, тем не менее, проследить наследование терминов одних кластеров другими, относящимися к следующему временному срезу, можно и на данных такого типа. Например, выявление аргументации, имеющей отношение к анализу мнений, графовый метод ее представления просматривается в кластерах 2 (2015), 1(2016), 2(2017), при этом можно заметить появление в методах 2016 г. позиционной контекстной классификации (context, stance classification).

Среди информативных вышедших из оборота терминов можно назвать следующие: wordnet, news article, large set, politic, original claim, trigram, adjective, adverb, main verb, class distribution, probability distribution (отсутствуют в 2016 г.); tree, baseline model, annotation guideline, discourse, elementary discourse unit, linguistic feature, controversial topic, semantic similarity, random forest, n-gram, fold cross validation, binary feature (отсутствуют в 2017 г.). Нельзя сказать, что вышедшие термины оказались менее актуальными, но они стали реже обсуждаться в данный момент времени.

Значительное обновление терминологии происходило в каждом временном периоде. В 2016 г. это: stance classification, baseline model, sentiment feature, n-gram, confusion matrix, persuasive essay, annotation guideline, elementary discourse unit, linguistic feature, fold cross validation, cohens kappa, f1 score, statistic, argumentation structure, argumentative microtext, controversial topic, parser, rebuttal, tree structure, main claim, svm, support vector machine;

2017 г.: ranking, supervised machine, web, logistic regression, pipeline, tf idf, wikipedia, word embedding, argumentative relation, argument scheme, argument unit.

На графе легко увидеть возникшие и исчезнувшие кластеры, содержащие практически новую лексику: 4-й и 5-й в 2016 г., один из которых в следующем году прекращает свое существование (5-й).

5. Заключение

Результатом проделанной работы является создание pipeline для проведения исследований по выявлению тематических кластеров ПО на основе коллекций текстов, а также отслеживанию изменений в их терминологическом составе в отдельные временные периоды.

Трансформацию терминологического состава кластеров предлагается анализировать с помощью ориентированных графов, построенных на основе критерия, который позволяет фиксировать наиболее значимые изменения. Терминологическая лексика выявленных тематических кластеров характеризует отдельные направления, в которых ведутся исследования, а трансформация терминологического состава кластеров во времени демонстрирует изменения, связанные с предпочтениями в выборе задач и методов.

Так, анализ предметной области Argument Mining выявил наличие неустоявшейся терминологии, стабильность в использовании методов на основе машинного обучения, а также конкурентоспособных методов на основе знаний. Отмечен переход от более простых моделей аргументации к более сложным. Данное исследование полезно при составлении обзоров ПО, выявления baseline, выборе актуальных методов и ресурсов для решения задач, связанных с анализом аргументации.

Планируется продолжить работы и провести расширенный эксперимент на коллекции текстов ПО Argument Mining, дополненной текстами других временных срезов.

Список литературы

1. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН, 2017. Т. 29, вып. 2. С. 161-200.
2. Callon M., Courtial, J.P. Laville. Co-word analysis as a tool for describing the network of interaction between basic and technological research: the case of polymer chemistry // Scientometrics. 1991. N 22. P. 155−205.
3. Small H. Tracking and predicting growth areas in science [Электронный ресурс]. URL: http://www.scimaps.org/exhibit/docs/small.pdf (дата обращения: 10.09.2019).
4. Van Eck N.J., Waltman L. Visualizing Bibliometric Networks [Электронный ресурс]. URL: https://link.springer.com/chapter/10.1007 (дата обращения: 10.09.2019).

5. Cobo M.J., López-Herrera A.G., Herrera-Viedma E., Herrera F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field [Электронный ресурс]. URL: https://www.sciencedirect.com/science/article/pii/S1751157710000891 (дата обращения: 10.09.2019).
6. Kandilas V., Uphum, S. P., Ungar L. H. Analyzing knowledge communities using foreground and background clusters [Электронный ресурс]. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.146.3141&rep= rep1&type=pdf (дата обращения: 12.09.2019).
7. VOSviewer Homepage, URL: http://www.vosviewer.com/, (дата обращения: 12.09.2019).
8. Waltman, L., Van Eck, N.J., & Noyons, E.C.M. A unified approach to mapping and clustering of bibliometric networks // Journal of Informetrics. 2010. N 4(4). P. 629-635.
9. Сивоголовко Е.В. Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании. 2011. № 4. С. 14–31.
10. Chou C.H., Su M.C., E. Lai. A new cluster validity measure and its application to image compression // Pattern Analysis and Applications. 2004.
11. Calinski R.B., Harabasz J. A dendrite method for cluster analysis // Comm. in Statistics. 1974.
12. Kaufman L., Rousseeuw P. Finding Groups in Data. An Introduction to Cluster Analysis. Wiley, 2005.